

# Energy Sharing for Multiple Sensor Nodes with Finite Buffers

Sindhu Padakandla, Prabuchandran K.J. and Shalabh Bhatnagar

Dept of Computer Science and Automation, Indian Institute of Science

E-Mail: {sindhupr, prabu.kj, shalabh}@csa.iisc.ernet.in

## Abstract

We consider the problem of finding optimal energy sharing policies that maximize the network performance of a system comprising of multiple sensor nodes and a single energy harvesting (EH) source. Sensor nodes periodically sense the random field and generate data, which is stored in the corresponding data queues. The EH source harnesses energy from ambient energy sources and the generated energy is stored in an energy buffer. Sensor nodes receive energy for data transmission from the EH source. The EH source has to efficiently share the stored energy among the nodes in order to minimize the long-run average delay in data transmission. We formulate the problem of energy sharing between the nodes in the framework of average cost infinite-horizon Markov decision processes (MDPs). We develop efficient energy sharing algorithms, namely Q-learning algorithm with exploration mechanisms based on the  $\epsilon$ -greedy method as well as upper confidence bound (UCB). We extend these algorithms by incorporating state and action space aggregation to tackle state-action space explosion in the MDP. We also develop a cross entropy based method that incorporates policy parameterization in order to find near optimal energy sharing policies. Through simulations, we show that our algorithms yield energy sharing policies that outperform the heuristic greedy method.

## Keywords:

Energy harvesting sensor nodes, energy sharing, Markov decision process, Q-learning, state aggregation.

## 1 Introduction

A sensor network is a group of independent sensor nodes, each of which senses the environment. Sensor networks find applications in weather and soil conditions monitoring, object tracking and structure monitoring. Each sensor node in the network senses the environment

and transmits the sensed data to a fusion node. The fusion node obtains data from several sensor nodes and carries out further processing.

In order to sense the environment and transmit data to the fusion node, nodes require energy and most often the nodes are equipped with pre-charged batteries for this purpose. However, as the nodes exhaust their battery power and stop sensing, the network performance degrades. The lifetime of the network is linked to the lifetimes of the individual nodes. Hence, the network becomes inoperable when a large number of nodes stop sensing. Thus, in a network with battery operated sensor nodes, the primary intention is to enhance the lifetime of the network, which may often lead to a compromise in the network performance. Many techniques have been proposed, which focus on improving lifetime of networks of sensor nodes. One of the more recent techniques which deals with this problem is the usage of energy harvesting to provide a perpetual source of energy for the nodes.

An energy harvesting (EH) sensor node replenishes the energy it consumes by harvesting energy from the environment (e.g., solar, wind power etc.) or other sources (e.g., body movements, finger strokes etc.) and converting into electrical energy. This way an EH node can be constantly powered through energy replenishment. So when compared to networks consisting of battery operated nodes, the long-term network performance metrics become appropriate. Thus, the goal pertaining to an EH sensor network is to reduce the average delay in data transmission. Even though an EH sensor node potentially has infinite amount of energy, yet the energy harvested is infrequently available as it is usually location and time dependent. Moreover the amount of energy replenished might be lower than the required amount. Therefore it is important to match the energy consumption with the amount of energy harvested in order to prevent energy starvation. This underlines the need for intelligently managing harvested energy to achieve the goal of good network performance.

A drawback associated with an EH sensor (node) is that it requires additional circuitry to harvest energy, which increases the cost of the node. A network which contains several such nodes is not economically viable. The cost of the network can be minimized if there exists a central EH source which harvests energy and shares the available energy among multiple sensor nodes in its vicinity. Such an architecture is incorporated in *motes*. A *mote* (Fig. 1) is a single unit on which sensors with different functionalities are arranged (see [13]). For instance, there could be pressure sensors, temperature sensors etc., in the same unit to make different sets of measurements simultaneously. Alternatively, the sensors could be of the same functionality but deployed together at different angles in order to have a 360° view of the entire sensing region.

Each of these sensors (within a unit) have their own data buffers and a common EH source feeds energy to each of the data queues. Usually, the EH source is a battery which is recharged by energy harvesting. The sensors in the mote are perpetually powered, but only if the energy harvested in the source is efficiently shared. Thus there is a need for a technique that dynamically allocates energy to each of the data buffers of individual sensors in order that the average queue lengths (or transmission delays) across the data buffers are minimized.

In this paper, we focus on the problem of developing algorithms that achieve efficient

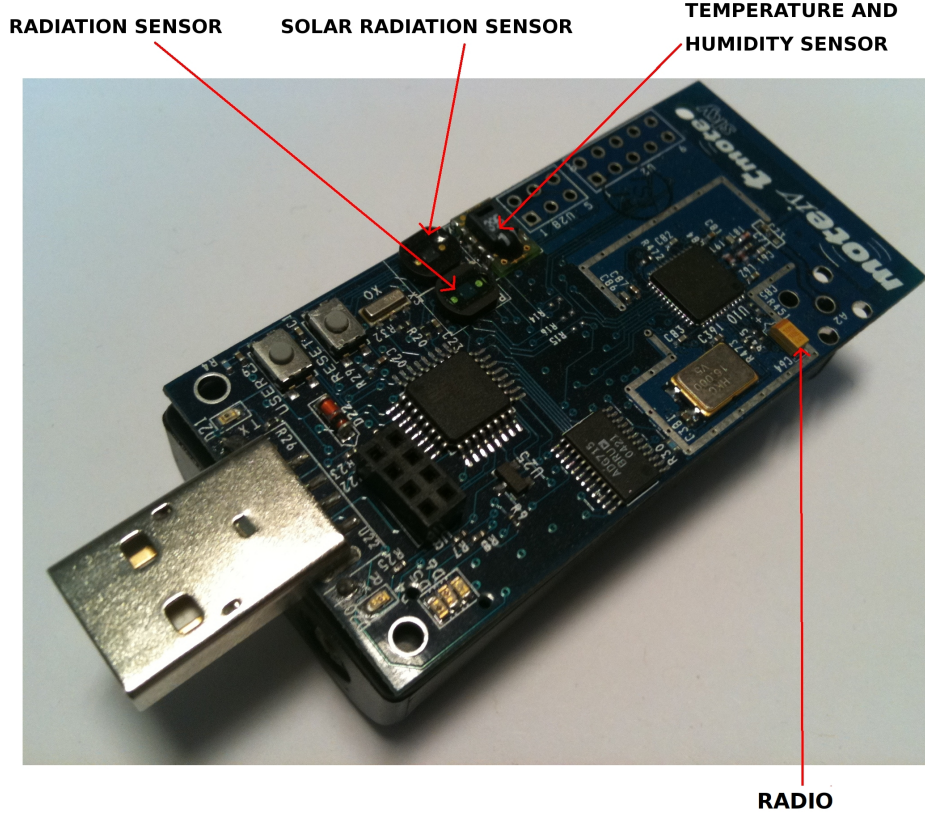


Figure 1: Mote with pressure, humidity and temperature sensors (Courtesy: Advanticsys Pvt Ltd. and UC Berkeley)

energy allocation in a system comprising of multiple sensor nodes with their own data buffers and a common EH source. Another scenario (that however we do not consider here) where our techniques are applicable is the case of downlink transmissions [22], where a base station (BS) maintains a separate data queue for each individual sensor node. The BS in question would also typically be powered by a single EH source, and again the problem would be to dynamically allocate the available energy to each one of the data queues. As suggested by a reviewer of the journal version of this paper, the above is equivalent to a communication setup with an energy harvesting transmitter and  $n$  receivers which are connected to the transmitter over orthogonal links and equal gain links. The transmitter employs  $n$  finite data buffers to store incoming data, intended for the  $n$  receivers and must optimally allocate its energy to transmit data intended for the  $n$  receivers.

We present learning algorithms for a controller which has to judiciously distribute the energy amongst the competing nodes. The controller decides on the amount of energy to be allocated to every node at every decision instant considering the amount of data waiting to be transmitted in each of the data queues. Thus the state of the system comprises of the

amount of data in each of the data queues along with the energy available in the source. Given the system state at an instant, the controller has to find out the best possible way to allocate energy to the individual nodes. The decided allocation has a bearing on the total amount of data transmitted at that instant as well as the amount of data that will be transmitted in the future. Our algorithms help the controller learn the optimal allocation for every state, one which reduces the buildup of data in the data buffers. In the algorithm we present, the controller systematically tries out several possible allocations feasible in a state, before learning the optimal allocation. This method is computationally efficient for small number of states. However it becomes computationally expensive when there are numerous states. We propose approximation algorithms to find the near-optimal allocation of energy in this scenario. In the following subsection, we survey literature on EH nodes and energy management policies employed in EH sensor networks.

## 1.1 Related Work

Optimizing energy usage in battery-powered sensors is addressed in [33, 34]. The problem of designing appropriate sensor schedules of sensor data transmission is discussed in [33]. A schedule of data transmission indicates when the battery-powered sensor transmits data. Transmitting data uses up energy, while not transmitting data results in error in estimation of parameters dependent on the sensor data. The authors in [33] consider battery-powered sensor nodes, each of which needs to minimize the energy utilized for data transmission. The estimation of parameters dependent on the sensor data may however involve error if the sensor does not transmit data for long periods of time. The objective in [33] is to find optimal periodic sensor schedules which minimize the estimation error at the fusion node and optimize energy usage.

In [34], the authors consider battery-powered sensors with two transmission power levels. The transmission power levels have different packet drop rates with the higher transmission power level having a lower packet drop rate. The sensor can choose one of the power levels for data transmission. It is assumed that the fusion node sends an acknowledgment (ACK or NACK) to the sensor node which indicates whether the data packet has been received or not. The objective in [34] is to minimize the average expected error in state estimation under energy constraint. At time  $k$ , based on the communication feedback the sensor knows whether the previous packets have been received by the fusion node or not. The problem of choosing the transmission power level is modeled as a MDP and the optimal schedule is shown to be stationary. The works [33, 34] consider the problem of efficient energy usage in battery powered sensors. The aspect of network performance is not considered in these. Our work deals with optimizing energy sharing in EH nodes where maximizing a network performance objective is the primary goal.

An early work in rechargeable sensors is [18]. The authors of [18] present a framework for the sensor network to adaptively learn the spatio-temporal characteristics of energy availability and provide algorithms to use this information for task sharing among nodes. In [17], the irregular and spatio-temporal characteristics of harvested energy are considered. The authors discuss the conditions for ensuring *energy-neutral* operation, i.e., using the energy

harvested at an appropriate rate such that the system continues to operate forever. Practical methods for a harvesting system to achieve energy-neutral operation are developed. Compared to [18, 17], we focus on minimizing the delay in data transmission from the nodes and also ensuring energy neutral operation.

The scenario of a single EH transmitter with limited battery capacity is considered in [41, 26]. In [26], the transmitter communicates in a fading channel, whereas in [41], no specific constraints on the channel are considered. The problem of finding the optimal transmission policy to maximize the short-term throughput of an EH transmitter is considered in [41]. Under the assumption of an increasing concave power-rate relationship, the short-term throughput maximizing transmission policy is identified. In [26], the transmitter gets channel state information and the node has to adaptively control the transmission rate. The objective is to maximize the throughput by a deadline and minimize the transmission completion time of a communication session. The authors in [26] develop an online algorithm which determines the transmit power at every instant by taking into account the amount of energy available and channel state.

The efficient usage of energy in a single EH node has been dealt with in some recent works [25, 36, 30, 46]. A channel and data queue aware sleep/active/listen mechanism in this direction is proposed in [25]. Listen mode turns off the transmitter, while sleep mode is activated if channel quality is bad. The node periodically enters the active mode. In the listen mode, the queue can build up resulting in packets being dropped. In the sleep mode, incoming packets are blocked. A bargaining game approach is used to balance the probabilities of packet drop and packets being blocked. The Nash equilibrium solution of the game controls the sleep/active mode duration and the amount of energy used.

The model proposed in [36, 30] considers a single EH sensor node with finite energy and data buffers. The authors assume that data sensed is independent across time instants and so is the energy harvested. The amount of data that can be transmitted using some specified energy is modeled using a *conversion function*. In [36], a linear conversion function is used and optimal energy management policies are provided for the same. These policies are throughput optimal and mean delay optimal in a low SNR regime. However, in the case of non-linear conversion function, [36] provides certain heuristic policies. In [30], a non-linear conversion function is used. The authors therein provide simulation-based learning algorithms for the energy management problem. These algorithms are model-free, i.e., do not require an explicit model of the system and the conversion function. Unlike [41, 26, 36, 25, 30], our work deals with multiple sensors sharing a common EH power source. The maximization objective is the delay in data transmission from the nodes. However, channel constraints are not addressed in our work.

Data packet scheduling problems in EH sensor networks are considered in [46] and [45]. It is assumed in [46] that a single EH node has separate data and energy queues, while the data sensed and energy harvested are random. The same assumption is made for each sensor in a two-sensor communication system considered in [45]. For simplicity it is assumed that all data bits have arrived in the queue and are ready for transmission, while the energy harvesting times and harvested energy amounts are known before the transmission

begins. In [46]([45]) the objective is to minimize the time by which all data packets from the node(s) are transmitted (to the fusion node). It is proposed to optimize this by controlling the transmission rate. The authors develop an algorithm to find the transmission rate at every instant, which optimizes the time to transmit the data packets. A two-user Gaussian interference channel with two EH sensor nodes and receivers is considered in [42]. This paper focuses on short-term sum throughput maximization of data transmitted from the two nodes before a given deadline. The authors provide generalized water-filling algorithms for the same. In contrast to the models developed in [46, 45, 42], our model assumes multiple sensors sharing a common energy source. The data and energy arrivals are uncertain and unknown. Moreover the problem we deal with has an infinite horizon, wherein the objective is to reduce the mean delay of data transmission from the nodes. We develop simulation based learning algorithms for this problem.

Cooperative wireless network settings are considered in [10, 15, 43]. Three different network settings with energy transfer between nodes are considered in [15]. Energy management policies which maximize the system throughput within a given duration are determined in all the three cases. A water-filling algorithm is developed which controls the flow of harvested energy over time and among the nodes. In [43], there exists an EH relay node and multiple other EH source nodes. The source nodes have infinite data buffer capacity. The relay node transfers data between the source and destination nodes. The source and relay nodes can transfer energy to one another. A sum rate maximization problem in this setting is solved. In [10], multiple pairs of sources and destinations communicate via an EH relay node. The EH relay node has a limited battery, which is recharged by wireless energy transfer from the source nodes. The EH relay node has to efficiently distribute the power obtained among the multiple users. The authors investigate four different power allocation strategies for outage performance (outage is an event in which data is lost due to lack of battery energy or transmission failures caused by channel fades). We do not consider energy cooperation between nodes in the sensor network. Moreover, we do not assume wireless energy transfer in our model.

A multi-user additive white Gaussian noise (AWGN) broadcast channel comprising of a single EH transmitter and  $M$  receivers is considered in [28]. The EH transmitter harvests energy from the environment and stores in a queue. The transmitter has  $M$  data queues, each of which stores data packets intended for a specific receiver. The data queues have fixed number of bits to be delivered to the receiver. The objective in [28] is to find a transmission policy that minimizes the time by which all the bits are transmitted to the receivers. An optimization problem is formulated and structural properties of the optimal policy are derived. In our work, we model energy sharing in multiple nodes when there is a single power source. We assume uncertain data and energy arrival processes. The objective is to minimize the average delay in data transmission from the nodes, when there is data arrival at every instant.

## 1.2 Our Contributions

- We consider the problem of efficient energy allocation in a system with multiple sensor nodes, each with its own data buffer, and a common EH source.
- We model the above problem as an infinite-horizon average cost Markov decision process (MDP) [4],[32] with an appropriate single-stage cost function. Our objective in the MDP setting is to minimize the long-run average delay in data transmission.
- We develop reinforcement learning algorithms which provide optimal energy sharing policies for the above problem. The learning procedure used does not need the system knowledge such as data and energy rates or cost structure and learns using the data obtained in an online manner.
- In order to deal with the dimensionality of the state space of the MDP, we present approximation algorithms. These algorithms find near-optimal energy distribution profiles when the state-action space of the MDP becomes unmanageable.
- We demonstrate through simulations that the policies obtained from our algorithm are better than the policies obtained from a heuristic greedy method and a combined nodes Q-learning algorithm (see Section 6).

## 1.3 Organization of the Paper

The rest of the paper is organized as follows. The next section describes the model, related notation and assumptions. Section 3 formulates the energy sharing problem as an MDP. Section 4 presents the RL algorithms used for solving the MDP. Section 5 highlights the need for approximate policies and gives a detailed explanation of the approximation algorithms we develop for the problem. Section 6 presents the simulation results of our algorithms. Section 7 provides the concluding remarks and possible future directions. Finally, an appendix at the end of the paper contains the proof of two results.

## 2 Model and Notation

We consider the problem of sharing the energy available in an energy harvesting source among multiple sensor nodes. We present a slotted, discrete-time, model (Fig. 2) for this problem. A sensor node in the network senses a random field and stores the sensed data in a finite data buffer of size  $D_{MAX}$ . In order to transmit the sensed data to a fusion (or central) node, the sensor node needs energy, which it obtains from an energy harvesting source. The energy harvesting source has an energy buffer of finite capacity  $E_{MAX}$ . The common EH source is an abstract entity in the model. It is generally a rechargeable battery which is replenished by random energy harvests. We assume fragmentation of data packets (fluid model) as in [36] and hence these will be treated as bit strings.

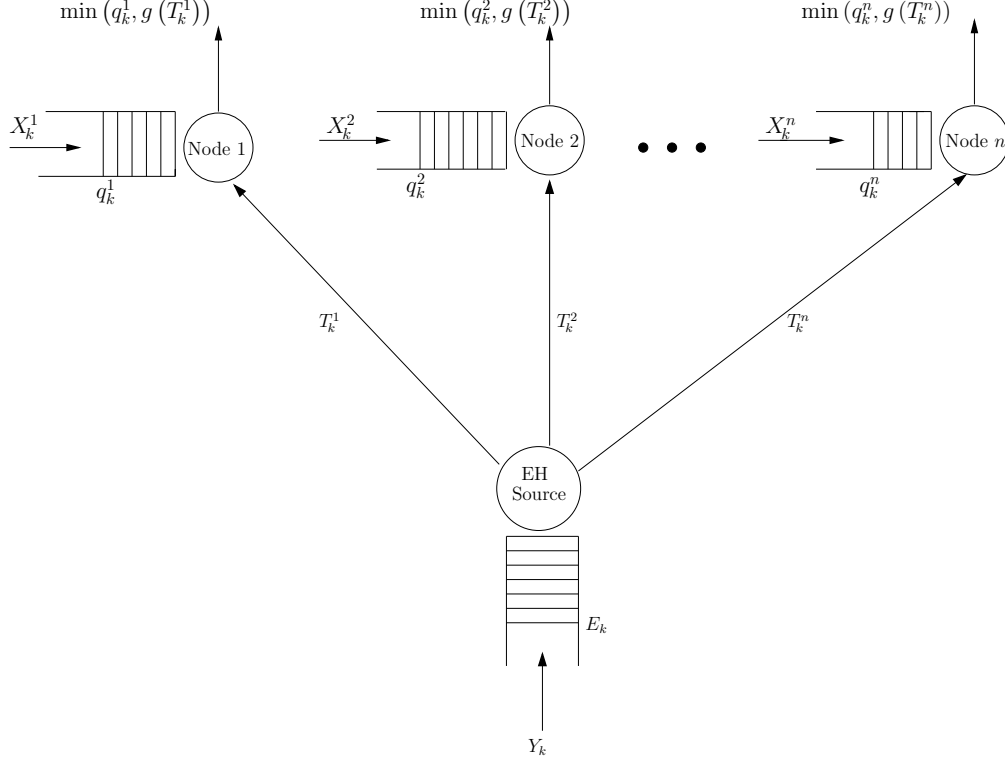


Figure 2: The System Model

Let  $q_k^i$  denote the data buffer level of node  $i$  and  $E_k$  be the energy buffer level at the beginning of slot  $k$ . Sensor node  $i$  generates  $X_k^i$  bits of data by sensing the random field. The source harvests  $Y_k$  units of energy. Based on the data queue levels  $(q_k^1, \dots, q_k^n)$  and the energy level  $E_k$ , the energy sharing controller decides upon the number of energy bits to be provided to every node. Let  $T_k^i$  units of energy be provided to node  $i$  in slot  $k$ . Using it, the node transmits  $g(T_k^i)$  bits of data. We have assumed the function  $g$  to be monotonically non-decreasing and concave as with other references ([36, 44, 16, 27, 14]). Note that the Shannon Channel capacity for Gaussian channels gives such a conversion function and in particular,

$$g(T_k) = \frac{1}{2} \log(1 + \beta T_k),$$

where  $\beta$  is a constant and  $\beta T_k$  gives the Signal-to-Noise (SNR) ratio. This is a non-decreasing concave function. We have assumed this form in the simulation experiments. However, our algorithms work regardless of the form of the conversion function and will learn the optimal energy sharing policy for any form of conversion function (see Remark 15).

It should be noted that we do not consider wireless energy transfer from the source node to the sensor nodes. Here we consider the source node to be a rechargeable battery which powers the nodes. The queue lengths in the data buffers evolve with time as follows:

$$q_{k+1}^i = (q_k^i - g(T_k^i))^+ + X_k^i \quad 1 \leq i \leq n, k \geq 0, \quad (1)$$



where  $(q_k^i - g(T_k^i))^+ = \max(q_k^i - g(T_k^i), 0)$  and the energy buffer queue length evolves as given below:

$$E_{k+1} = \left( E_k - \sum_{i=1}^n T_k^i \right) + Y_k, \quad 1 \leq i \leq n, k \geq 0, \quad (2)$$

where  $\sum_{i=1}^n T_k^i \leq E_k$ .

**Assumption 1.** *The generated data rates at time  $k + 1$ ,  $X_{k+1} \triangleq (X_{k+1}^1, X_{k+1}^2, \dots, X_{k+1}^n)$  where  $n$  denotes the number of sensors in a node, evolves as a jointly Markov process, i.e.,*

$$X_{k+1} = f^1(X_k, W_k), \quad k \geq 0 \quad (3)$$

where  $f^1$  is some arbitrary vector valued function with  $n$  components and  $\{W_k, k \geq 1\}$  is a noise sequence with probability distribution  $P(W_k | X_k)$  depending on  $X_k$ . Thus, the generated data  $\{X_k, k \geq 0\}$  is both spatially and temporally correlated. Moreover, the sequence  $X_k^i, k \geq 0$  satisfies  $\sup_{k \geq 0} \mathbb{E}[X_k^i] \leq r < \infty$ . Further, the energy arrival process evolves as:

$$Y_{k+1} = f^2(Y_k, V_k), \quad k \geq 0, \quad (4)$$

where  $f^2$  is some scalar valued function and  $\{V_k, k \geq 1\}$  is the noise sequence with probability distribution  $P(V_k | Y_k)$  depending on  $Y_k$ .

**Remark 1.** *Assumption 1 is general enough to cover most of the stochastic models for the data and energy arrivals. A special case of Assumption 1 is to consider that for any  $k \geq 0$  and  $1 \leq i \leq n$ ,  $X_k^i$  is independent of  $X_{k-1}^i, X_{k-2}^i, \dots, X_1^i, X_0^i$  and the given sequence  $\{X_k^i\}_{k \geq 0}$  for a given  $i \in \{1, \dots, n\}$  is identically distributed. Similarly, for any  $k \geq 0$ ,  $Y_k$  is independent of  $Y_{k-1}, Y_{k-2}, \dots, Y_1, Y_0$  and the sequence  $\{Y_k\}$  is identically distributed. In Section 6 we show results of experiments where the above i.i.d setting as well as a more general setting as described earlier are shown.*

### 3 Energy Sharing Problem as an MDP

A Markov decision process (MDP) is a tuple of states, actions, transition probabilities and single-stage costs. Given that the MDP is in a certain state, and an action is chosen by the controller, the MDP moves to a ‘next’ state according to the prescribed transition probabilities. The objective of the controller is to select a sequence of actions as a function of the states in order to minimize a given long-term objective (cost). We formulate the energy sharing problem in the MDP setting using the long-run average cost criterion. The MDP formulation requires that we identify the states, actions and the cost structure for the problem, which is described next.

The state  $s_k$  is a tuple comprising of the data buffer level of all sensor nodes, the level of the energy buffer in the source, the data and energy arrivals in the past. Note that for

$1 \leq i \leq n$ ,  $q_k^i \in \{0, 1, \dots, D_{MAX}\}$ . Similarly  $E_k \in \{0, 1, \dots, E_{MAX}\}$ . Thus in stage  $k$ , in the context of Assumption 1, state  $s_k = (q_k^1, q_k^2, \dots, q_k^n, E_k, X_{k-1}, Y_{k-1})$ . However, when we assume that for  $1 \leq i \leq n$ ,  $\{X_k^i\}$  and  $\{Y_k\}$  are i.i.d (as in Remark 1), then the state tuple simplifies to  $s_k = (q_k^1, q_k^2, \dots, q_k^n, E_k)$ .

The set of all states is the state-space, which is denoted by  $S$ . Similarly  $A$  denotes the action-space, which is the set of all actions. The set of feasible actions in a state  $s_k$  is denoted by  $A(s_k)$ . A deterministic policy  $\pi = \{T_k, k \geq 0\}$  is a sequence of maps such that at time  $k$  when state  $s_k = (q_k^1, \dots, q_k^n, E_k, X_{k-1}, Y_{k-1})$ , i.e., when there are  $q_k^j$  units of data at node  $j$ ,  $1 \leq j \leq n$  and  $E_k$  bits of energy in the source,  $X_k$  is the data arrival vector and  $Y_k$  is the energy harvested at time  $k$ , then  $T_k(s_k) = (T_k^1(s_k), T_k^2(s_k), \dots, T_k^n(s_k))$  gives the number of energy bits to be given to each node at time  $k$  (i.e., it gives the energy split). Thus the action to be taken in state  $s_k$  is given by  $T_k(s_k) \in A(s_k)$ . A deterministic policy which does not change with time is referred to as a stationary deterministic policy (SDP). We denote such a policy  $\pi$  as  $\pi = (T, T, \dots)$ , where  $T(s_k)$  is the action chosen in state  $s_k$ . We set the single-stage cost  $\tilde{c}(s_k, T(s_k))$  as a sum of the number of bits in the data buffers. Thus,

$$\tilde{c}(s_k, T(s_k)) = \sum_{i=1}^n q_k^i. \quad (5)$$

**Remark 2.** In order to formulate the energy sharing problem in the framework of MDP, we require the state sequence  $\{s_k = (q_k^1, q_k^2, \dots, q_k^n, E_k)\}_{k \geq 0}$  under a given policy to be a Markov chain, i.e.,

$$P(s_{k+1} \mid s_k, s_{k-1}, \dots, s_0, \pi) = P(s_{k+1} \mid s_k, \pi).$$

We have generalized the assumption on  $\{X_k^i, 1 \leq i \leq n\}_{k \geq 0}$  and  $\{Y_k\}_{k \geq 0}$  and consider jointly Markov data arrival and Markovian energy arrival processes. Remark 1 applies to the i.i.d case. If we assume the data arrivals  $\{X_k^i\}_{k \geq 0}$  for a fixed  $i \in \{1, 2, \dots, n\}$  and the energy arrivals  $\{Y_k\}_{k \geq 0}$  are i.i.d, then the Markov assumption can be seen to be easily satisfied.

The Markov property for the state evolution  $\{s_k\}_{k \geq 0}$  is necessary as we can only search for policies based only on the present state of the system. Otherwise, the policies will be based on the entire history. The search for optimal policies in the space of history based policies is a computationally infeasible task.

In the general case where  $\{X_k\}$  is jointly Markov, note that the state sequence  $\{s_k\}_{k \geq 0}$  under a given policy will not be a Markov chain. Now consider the augmented state  $\bar{s}_k \triangleq \begin{pmatrix} s_k \\ X_{k-1} \\ Y_{k-1} \end{pmatrix}$ . Now, under a given policy  $\pi = (T, \dots, T)$ , the state evolution can be described as

$$\begin{pmatrix} q_{k+1}^1 \\ \vdots \\ q_{k+1}^n \\ E_{k+1} \\ X_k \\ Y_k \end{pmatrix} = \begin{pmatrix} (q_k^1 - g(T^1(s_k)))^+ + X_k^1 \\ \vdots \\ (q_k^n - g(T^n(s_k)))^+ + X_k^n \\ (E_k - \sum_{i=1}^n T^i(s_k)) + Y_k, \\ f^1(X_{k-1}, W_{k-1}) \\ f^2(Y_{k-1}, V_{k-1}) \end{pmatrix}. \quad (6)$$

This can be written as  $\bar{s}_{k+1} = h(\bar{s}_k, T(\bar{s}_k), W_{k-1}, V_{k-1})$  for suitable vector valued function  $h$ . This is the standard description for the state evolution for an MDP (see Chapter 1 in [3]). Since the probability distribution of the noise  $W_{k-1}$  ( $V_{k-1}$ ) depends only on  $X_{k-1}$  ( $Y_{k-1}$ ), the augmented state sequence  $\bar{s}_k = \{(s_k, X_{k-1}, Y_{k-1})\}_{k \geq 0}$  forms a Markov chain. This facilitates search for policies only based on the present augmented state.

**Remark 3.** The sensor node may generate data as packets, but in the model we allow for arbitrary fragmentation of data during transmission. Hence packet boundaries are no longer relevant and we consider bit strings. This is the fluid model as described in [11]. The data is considered to be stored in the data buffers as bit strings and hence the data buffer levels are discrete. The fluid model assumption (data discretization) has been made in [36, 14, 46]. For energy harvesting we consider energy discretization. Energy discretization implies that we have assumed that discrete levels of energy are harvested and stored in the queue. Energy discretization has been considered in some previous works [2, 36]. Owing to these assumptions on data generation and energy harvesting, the state space is discrete and finite.

The long-run average cost of an SDP  $\pi$  is given by

$$\tilde{\lambda}^\pi = \lim_{m \rightarrow \infty} \mathbb{E} \left[ \frac{1}{m} \sum_{k=0}^{m-1} \tilde{c}(s_k, T(s_k)) \right]. \quad (7)$$

In contrast, a stationary randomized policy (SRP) is a sequence of maps  $\varphi = \{\psi, \psi, \dots\}$  such that for a state  $s_k$ ,  $\psi(s_k, \cdot)$  is a probability distribution over the set of feasible actions in state  $s_k$ . Such a policy does not change with time. The single-stage cost  $\tilde{d}(s_k)$  of an SRP  $\varphi$  is given by

$$\tilde{d}(s_k) = \sum_{a \in A(s_k)} \psi(s_k, a) \tilde{c}(s_k, a), \quad (8)$$

where  $a$  gives the energy split in state  $s_k$ . The long-run average cost of an SRP  $\varphi$  is

$$\tilde{\lambda}^\varphi = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{k=0}^{m-1} \tilde{d}(s_k). \quad (9)$$

We observe that the term  $q_k^i$  in (5) does not include the effect of action explicitly. Hence we modify the cost function to include the effect of the action taken explicitly into the cost function. In order to enable reformulation of the average cost objective in the modified form, we prove the following lemma. Define

$$\lambda^\pi = \lim_{m \rightarrow \infty} \mathbb{E} \left[ \frac{1}{m} \sum_{k=0}^{m-1} \sum_{i=1}^n (q_k^i - g(T^i(s_k)))^+ \right]. \quad (10)$$

**Lemma 1.** Let  $q_k^i$ ,  $1 \leq i \leq n$ ,  $T^i(s_k)$ ,  $1 \leq i \leq n$  and  $g$  be as before and let  $\mathbb{E}[X^i]$ ,  $1 \leq i \leq n$  denote the mean of the i.i.d random variables  $X^i$ ,  $1 \leq i \leq n$ . Then

$$\lambda^\pi = \tilde{\lambda}^\pi - \sum_{i=1}^n \mathbb{E}[X^i]$$

for all policies  $\pi$ .

*Proof.* Using state evolution equations (1)-(2),

$$\begin{aligned}
& \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} \sum_{i=1}^n (q_k^i - g(T^i(s_k)))^+ \right] \\
&= \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} \sum_{i=1}^n (q_{k+1}^i - X_k^i) \right] \\
&= \lim_{m \rightarrow \infty} E \left[ \sum_{i=1}^n \frac{1}{m} \sum_{k=0}^{m-1} (q_{k+1}^i - X_k^i) \right] \\
&= \sum_{i=1}^n \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} (q_{k+1}^i - X_k^i) \right] \\
&= \sum_{i=1}^n \left\{ \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} q_{k+1}^i \right] - \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} X_k^i \right] \right\} \\
&= \sum_{i=1}^n \left\{ \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \left( \sum_{k=0}^{m-1} q_k^i + q_m^i - q_0^i \right) \right] - \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} X_k^i \right] \right\} \\
&= \sum_{i=1}^n \left\{ \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \left( \sum_{k=0}^{m-1} q_k^i + q_m^i - q_0^i \right) \right] - E[X^i] \right\} \\
&= \sum_{i=1}^n \left\{ \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} q_k^i \right] + \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} (q_m^i - q_0^i) \right] - E[X^i] \right\} \\
&= \sum_{i=1}^n \left\{ \lim_{m \rightarrow \infty} E \left[ \frac{1}{m} \sum_{k=0}^{m-1} q_k^i \right] \right\} - \sum_{i=1}^n E[X^i] \\
&= \tilde{\lambda}^\pi - \sum_{i=1}^n E[X^i].
\end{aligned}$$

The second last equality above follows from the fact that  $\lim_{m \rightarrow \infty} E \left[ \frac{1}{m} (q_m^i - q_0^i) \right] = 0$ . The claim follows.  $\square$

The linear relationship between  $\tilde{\lambda}^\pi$  and  $\lambda^\pi$  enables us to define the new single-stage cost function as:

$$c(s_k, T_k) = \sum_{i=1}^n (q_k^i - g(T^i(s_k)))^+. \quad (11)$$

With this single-stage cost function, the long-run average cost of an SDP  $\pi$  is given by

$$\lambda^\pi = \lim_{m \rightarrow \infty} \mathbb{E} \left[ \frac{1}{m} \sum_{k=0}^{m-1} c(s_k, T(s_k)) \right]. \quad (12)$$

The single-stage cost  $d(s_k)$  of an SRP  $\varphi$  is given by

$$d(s_k) = \sum_{a \in A(s_k)} \psi(s_k, a) c(s_k, a), \quad (13)$$

where  $a$  gives the energy split. The long-run average cost of an SRP  $\varphi$  is

$$\lambda^\varphi = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{k=0}^{m-1} d(s_k). \quad (14)$$

It can be inferred from Lemma 1 that a policy which minimizes the average cost in (11) (or (14)) will also minimize the average cost given by (7) (or (9)). In this paper we are interested in finding stationary policies (deterministic or randomized) which optimally share the energy among a set of nodes. Therefore our aim is to find policies which minimize the average cost per step, when the single-stage cost is given by (11).

Any stationary optimal policy minimizes the average cost of the system over all policies. Let  $\pi^*$  be an optimal policy and  $\Pi$  be the set of all policies. The average cost of policy  $\pi^*$  is denoted  $\lambda^*$ . Then

$$\lambda^* = \inf_{\pi \in \Pi} \lambda^\pi.$$

The policy corresponding to the above average cost minimizes the sum of (data) queue lengths of all nodes. By Little's law, under stationarity, the average sum of data queue lengths at the sensor nodes is proportional to the average waiting time or delay of the arrivals (bits). Hence an average cost optimal policy minimizes the stationary mean delay as well.

The class of stationary deterministic policies is contained in the class of stationary randomized policies and in the system we consider, an optimal policy is known to exist in the class of stationary deterministic policies. We provide an algorithm which finds an optimal SDP. The algorithm is computationally efficient for small state and action spaces. However for large state-action spaces, the algorithm computations are expensive. To mitigate this problem, we provide approximation algorithms which find near-optimal stationary policies for the system. These algorithms are described in the following sections.

## 4 Energy Sharing Algorithms

### 4.1 Background

Consider an optimal SDP  $\pi^*$  for the energy sharing MDP. Then  $\lambda^*$  corresponds to the average cost of the policy  $\pi^*$ . Suppose  $i_r$  is a reference state in the MDP. For any state  $i \in S$ , let  $h^*(i)$  be the relative (or the differential) cost defined as the minimum of the difference between the expected cost to reach state  $i_r$  from  $i$  and the expected cost incurred if the cost per stage was  $\lambda^*$ . The quantities  $\lambda^*$  and  $h^*(i), i \in S$  satisfy the Bellman Equation:

$$\lambda^* + h^*(i) = \min_{a \in A(i)} \left( c(i, a) + \sum_{j \in S} p(i, a, j) h^*(j) \right), \quad (15)$$

where  $p(i, a, j)$  is the probability that the system will move from state  $i$  to state  $j$  under action  $a$ . We denote by  $Q^*(i, a)$ , the optimal differential cost of any feasible state-action tuple  $(i, a)$  as follows:

$$Q^*(i, a) = c(i, a) + \sum_{j \in S} p(i, a, j) h^*(j). \quad (16)$$

Equation (15) can now be rewritten as

$$\lambda^* + h^*(i) = \min_{a \in A(i)} Q^*(i, a), \quad \forall i \in S \quad (17)$$

or alternately

$$h^*(i) = \min_{a \in A(i)} Q^*(i, a) - \lambda^*, \quad \forall i \in S. \quad (18)$$

Plugging (18) into (16), one obtains

$$Q^*(i, a) = c(i, a) + \sum_{j \in S} p(i, a, j) \left[ \min_{b \in A(j)} Q^*(j, b) - \lambda^* \right] \quad (19)$$

or

$$\lambda^* + Q^*(i, a) = c(i, a) + \sum_{j \in S} p(i, a, j) \min_{b \in A(j)} Q^*(j, b), \quad \forall i \in S, \forall a \in A(i). \quad (20)$$

Equation (20) is also referred to as the Q-Bellman equation. The important thing to note is that whereas the Bellman equation (15) is not directly amenable to stochastic approximation, the Q-Bellman equation (20) is; because of the fact that the minimization operation in (20) is inside the conditional expectation unlike (15) (where it is outside of it). If the transition probabilities and the cost structure of the system model are known, then (20) can be solved using dynamic programming techniques [40]. When the system model is not known (as in the problem we study), the Q-learning algorithm can be used to obtain optimal policies. This learning algorithm solves (20) in an online manner using simulation to obtain an optimal policy. It is described in the following subsection.

## 4.2 Relative Value Iteration based Q-Learning

Q-learning is a stochastic iterative, simulation-based algorithm that aims to find the  $Q^*(i, a)$  values for all feasible state-action pairs  $(i, a)$ . It is a model-free learning algorithm and proceeds by assuming that the transition probabilities  $p(i, a, j)$  are unknown. Initially Q-values for all state-action pairs are set to zero, i.e.,  $Q_0(i, a) = 0, \forall i \in S, a \in A(i)$ . Then  $\forall k \geq 0$ , the Q-learning update [1] for a state-action pair visited during simulation is carried out as follows:

$$Q_{k+1}(i, a) = (1 - \alpha(k))Q_k(i, a) + \alpha(k) \left( c(i, a) + \min_{b \in A(j)} Q_k(j, b) - \min_{u \in A(i_r)} Q_k(i_r, u) \right), \quad (21)$$

where  $i$  is the current state at decision time  $k$  and  $i_r$  is the reference state. The action in state  $i$  is selected using one of the exploration mechanisms described below. State  $j$  corresponds

to the ‘next’ state that is obtained from simulation when the action  $a$  is selected in state  $i$ . Also,  $\alpha(k)$ ,  $k \geq 0$  is a given step-size sequence such that  $\alpha(k) > 0, \forall k \geq 0$  and satisfies the following conditions:

$$\sum_k \alpha(k) = \infty \text{ and } \sum_k \alpha^2(k) < \infty.$$

Let  $t(k) = \sum_{i=0}^{k-1} \alpha(i)$ ,  $k \geq 1$ , with  $t(0) = 0$ . Then,  $t(k)$ ,  $k \geq 0$  corresponds to the “timescale” of the algorithm’s updates. The first condition above ensures that  $t(k) \rightarrow \infty$  as  $k \rightarrow \infty$ . This ensures that the algorithm does not converge prematurely. The second condition makes sure that the noise asymptotically vanishes. These conditions on step sizes guarantee the convergence of Q-learning to the optimal state-action value function, see [1] for a proof of convergence of the algorithm. The update (21) is carried out for the state-action pairs visited during simulation. The exploration mechanisms we employ are as follows:

1.  $\epsilon$ -greedy: In the energy sharing problem, the number of actions feasible in every state is finite. Hence there exists an action  $a_m$  for state  $i$  such that  $Q_k(i, a_m) \leq Q_k(i, a')$ ,  $\forall a' \in A(i)$ ,  $\forall k \geq 0$ . We choose  $\epsilon \in (0, 1)$ . In state  $i$ , action  $a_m$  is picked with probability  $1 - \epsilon$ , while any other action is picked with probability  $\epsilon$ .
2. UCB Exploration: Let  $N_i(k)$  be the number of times state  $i$  is visited until time  $k$ . Similarly let  $N_{i,a}(k)$  be the number of times action  $a$  is picked in state  $i$  upto time  $k$ . The Q-value of state-action pair  $(i, a)$  at time  $k$  is  $Q_k(i, a)$ . When the state  $i$  is encountered at time  $k$ , the action for this state is picked according to the following rule:

$$a' = \arg \max_{a \in A(i)} \left( -Q_k(i, a) + \beta \sqrt{\frac{\ln N_i(k)}{N_{i,a}(k)}} \right), \quad (22)$$

where  $\beta$  is a constant. The first term on the right hand side gives preference to an action that has yielded good performance in the past visits to state  $i$ , while the second term gives preference to actions that have not been tried out many times so far, relative to  $\ln N_i(k)$ .

**Remark 4.** *The convergence rates for the discounted Q-learning have been studied in [39, 19, 12]. The finite-time bounds to reach an  $\epsilon$ -optimal policy by following the Q-learning rule are given in [39, 19, 12]. In the Q-learning algorithm, to explore the value of different states and actions, one needs to visit each state-action pair infinitely often. However, in practice, depending on the size of the state-action space, we need to simulate the Q-learning algorithm so that each state-action pair is visited a sufficient number of times. In our experiments for the case of two sensor nodes, the size of the state space is of the order of  $10^5$  and we ran our algorithm for  $10^8$  iterations.*

Once we determine  $Q^*(i, a)$  for all state-action pairs, we can obtain the optimal action for a state  $i$  by choosing the action that minimizes  $Q^*(i, a)$ . So

$$a^* = \arg \min_{a \in A(i)} Q^*(i, a). \quad (23)$$

It should be noted that the Q-learning algorithm does not need knowledge of the cost structure and transition probabilities, and it learns an optimal policy by interacting with the system.

## 5 Approximation Algorithms

The learning algorithm described in Section 4 is an iterative stochastic algorithm that learns the optimal energy split. This method requires that the  $Q(s, a)$  values be stored for all  $(s, a)$  tuples. The values of  $Q(s, a)$  for each  $(s, a)$  tuple are updated in (21) over a number of iterations using adequate exploration. These updates play a key role in finding the optimal control for a given state. Nevertheless for large state-action spaces these computations are expensive as every lookup operation and updation require memory access. For example, if there are two nodes sharing energy and buffer sizes are  $E_{MAX} = D_{MAX} = 30$ , then the number of  $(s, a)$  tuples would be of the order  $10^6$ , which demands enormous amount of computation time and memory space. This condition is exacerbated when the number of nodes that share energy increases. For instance, in the case of four nodes sharing energy with  $E_{MAX} = D_{MAX} = 30$ , we have  $|S \times A| \approx 30^9$ . Thus, we have a scenario where the state-action space can be extremely large.

To mitigate this problem, we propose two algorithms that are both based on certain threshold features. Both algorithms tackle the curse of dimensionality, by reducing the computational complexity. We describe below our threshold based features, following which we describe our algorithms.

### 5.1 Threshold based Features

The fundamental idea of threshold based features is to cluster states in a particular manner, based on the properties of the differential value functions. The following proposition proves the monotonicity property of the differential value functions for the scenario where there is a single node and an EH source. This simple scenario is considered for the sake of clarity in the proof.

**Proposition 1.** *Let  $H^*(q, E)$  be the differential value of state  $(q, E)$ . Let  $q < q^L \leq D_{MAX}$  and  $E_{MAX} \geq E^L > E$ , respectively. Then,*

$$H^*(q, E) \leq H^*(q^L, E), \quad (24)$$

$$H^*(q, E) \geq H^*(q, E^L). \quad (25)$$

*Proof.* Let  $J(s)$  be the total cost incurred when starting from state  $s$ . Define the Bellman operator  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  as

$$(LJ)(s) = \min_{T \in A(s)} (c(s, T) + \mathbb{E}[J(s')]),$$

where  $s'$  corresponds to the next state after  $s$  and  $T$  corresponds to the action taken in state  $s$ . As noted in Section 5.1, we show the proof for a single node and EH source. The proof



can be easily generalized to multiple nodes. Thus the state  $s$  corresponds to the tuple  $(q, E)$ . Hence the above equation can be rewritten as

$$(LJ)(q, E) = \min_{T \in A(s)} \left( c(q, E, T) + \mathbb{E}[J(q', E')] \right), \quad \forall (q, E) \in S.$$

We consider the application of the operator  $L$  on the differential cost function  $H(\cdot)$ . We set out to prove this proposition using the relative value iteration scheme (see [5]). For this, we set a reference state  $r \triangleq (q_r, E_r) \in S$ . The cost function in our case is  $(q - g(T))^+$ . Initially the differential value function has value zero for all states  $(q, E) \in S$ , i.e.,  $H(q, E) = 0$ ,  $\forall (q, E) \in S$ . Then for some arbitrary  $(q, E) \in S$  we have

$$\begin{aligned} LH(q, E) &= \min_{T \in A(q, E)} \left( (q - g(T))^+ + \mathbb{E}[H(q', E')] \right) - LH(q_r, E_r) \\ &= \min_{T \in A(q, E)} ((q - g(T))^+) - LH(q_r, E_r) \end{aligned}$$

since  $H(q', E') = 0$ ,  $\forall (q', E') \in S$ . Let  $T_m$  be the value of  $T$  achieving the minimum in the first term of RHS. Then

$$LH(q, E) = (q - g(T_m))^+ - LH(q_r, E_r).$$

Now consider the differential value of state  $q^L$  where  $q^L > q$ . Thus, consider

$$\begin{aligned} LH(q^L, E) &= \min_{T \in A(q^L, E)} \left( (q^L - g(T))^+ + \mathbb{E}[H(q', E')] \right) - LH(q_r, E_r) \\ &= \min_{T \in A(q^L, E)} ((q^L - g(T))^+) - LH(q_r, E_r) \\ &= (q^L - g(T_L))^+ - LH(q_r, E_r), \end{aligned}$$

where  $T_L$  is the value of  $T$  for which the minimum of the expression  $(q^L - g(T))^+$ , in the above equations, is achieved. We have

$$\begin{aligned} LH(q, E) &= (q - g(T_m)) - LH(q_r, E_r) \\ &\leq (q - g(T_L)) - LH(q_r, E_r) \\ &\leq (q^L - g(T_L)) - LH(q_r, E_r) \\ &= LH(q^L, E). \end{aligned} \tag{26}$$

We have  $H(q^L, E) \geq H(q, E)$  since these values are initialized to zero and from (26),  $LH(q^L, E) \geq LH(q, E)$ . Now consider the differential value function of the state  $(q, E^L)$  where  $E^L > E$ .

$$\begin{aligned} LH(q, E^L) &= \min_{T \in A(q, E^L)} \left( (q - g(T))^+ + \mathbb{E}[H(q', E')] \right) - LH(q_r, E_r) \\ &= \min_{T \in A(q, E^L)} ((q - g(T))^+) - LH(q_r, E_r) \end{aligned}$$

$$= (q - g(T_E))^+ - L H(q_r, E_r),$$

where  $T_E$  is the value of  $T$  for which the minimum of the expression  $(q - g(T))^+$ , in the above equations, is achieved. We have,

$$\begin{aligned} L H(q, E^L) &= (q - g(T_E))^+ - L H(q_r, E_r) \\ &\leq (q - g(T_m))^+ - L H(q_r, E_r) \\ &\leq L H(q, E). \end{aligned} \tag{27}$$

Since  $H(q, E^L)$ ,  $H(q, E)$  are initialized to zero, we have  $H(q, E^L) \leq H(q, E)$  and from (27),  $L H(q, E^L) \leq L H(q, E)$ . We prove the following statements using mathematical induction:

$$\begin{aligned} L^k H(q, E) &\leq L^k H(q^L, E) \quad \forall k \geq 0, \\ L^k H(q, E) &\geq L^k H(q, E^L) \quad \forall k \geq 0. \end{aligned}$$

We have seen above that the two statements are true for both  $k = 0$  and  $k = 1$ , respectively. Lets consider the first statement and assume that the statement holds for some  $k$ . We then prove that it holds for  $(k + 1)$ . Consider

$$L^{k+1} H(q, E) = \min_{T \in A(q, E)} \left( (q - g(T))^+ + \mathbb{E}[L^k H(q', E')] \right) - L^k H(q_r, E_r).$$

Assume  $T_m$  is the value of  $T$  at which the minimum of  $((q - g(T))^+ + \mathbb{E}[L^k H(q', E')])$  is attained. Then,

$$L^{k+1} H(q, E) = ((q - g(T_m))^+ + \mathbb{E}[L^k H(q - g(T_m) + x, E - T_m + y)]) - L^k H(q_r, E_r),$$

where  $x, y$  are obtained from independent random distributions. Similarly, we get

$$L^{k+1} H(q^L, E) = ((q^L - g(T_L))^+ + \mathbb{E}[L^k H(q^L - g(T_L) + x, E - T_L + y)]) - L^k H(q_r, E_r),$$

where  $T_L$  is the value of  $T$  for which the minimum in the expression  $((q^L - g(T))^+ + \mathbb{E}[L^k H(q', E')])$  is achieved.

$$\begin{aligned} L^{k+1} H(q, E) &= ((q - g(T_m))^+ + \mathbb{E}[L^k H(q - g(T_m) + x, E - T_m + y)]) - L^k H(q_r, E_r) \\ &\leq ((q - g(T_L))^+ + \mathbb{E}[L^k H(q - g(T_L) + x, E - T_L + y)]) - L^k H(q_r, E_r) \\ &\leq ((q^L - g(T_L))^+ + \mathbb{E}[L^k H(q^L - g(T_L) + x, E - T_L + y)]) - L^k H(q_r, E_r), \end{aligned}$$

since the property holds true for  $L^k H$ , i.e.,  $L^k H(q, E) \leq L^k H(q^L, E)$ . Thus,

$$\begin{aligned} L^{k+1} H(q, E) &\leq ((q^L - g(T_L))^+ + \mathbb{E}[L^k H(q^L - g(T_L) + x, E - T_L + y)]) - L^k H(q_r, E_r) \\ &= L^{k+1} H(q^L, E). \end{aligned}$$

Hence,

$$L^k H(q, E) \leq L^k H(q^L, E) \quad \forall k \geq 0. \tag{28}$$

Similarly we get,

$$\begin{aligned}
L^{k+1}H(q, E^L) &= ((q - g(T_E))^+ + \mathbb{E}[L^k H(q - g(T_E) + x, E^L - T_E + y)]) - L^k H(q_r, E_r) \\
&\leq ((q - g(T_m))^+ + \mathbb{E}[L^k H(q - g(T_m) + x, E^L - T_m + y)]) - L^k H(q_r, E_r) \\
&\leq ((q - g(T_m))^+ + \mathbb{E}[L^k H(q - g(T_m) + x, E - T_m + y)]) - L^k H(q_r, E_r) \\
&= L^{k+1}H(q, E),
\end{aligned}$$

hence by mathematical induction on  $k$  we get,

$$L^k H(q, E^L) \leq L^k H(q, E) \quad \forall k \geq 0. \quad (29)$$

As a consequence of the relative value iteration scheme ([32]), when  $k \rightarrow \infty$ ,  $L^k H \rightarrow H^*$  with  $H^*(q_r, E_r) = \lambda^*$ . Thus, from (28) and (29) as  $k \rightarrow \infty$ , we obtain

$$H^*(q, E) \leq H^*(q^L, E)$$

$$H^*(q, E) \geq H^*(q, E^L).$$

The claim now follows.  $\square$

Proposition 1 can be easily generalized to multiple nodes in the following manner. Suppose there are  $n$  nodes and one EH source. Let  $s = (q^1, \dots, q^j, \dots, q^n, E)$  and  $s' = (q^1, \dots, q_L^j, \dots, q^n, E)$ , where  $q_L^j > q^j$ . The states  $s$  and  $s'$  differ only in the data buffer queue lengths of node  $j$ , while the data buffer queue lengths of other nodes remain the same and so does the energy buffer level. Then it can be observed that  $H^*(q^1, \dots, q^j, \dots, q^n, E) \leq H^*(q^1, \dots, q_L^j, \dots, q^n, E)$ . In a similar manner, let state  $s'' = (q^1, q^2, \dots, q^n, E^L)$  and  $E^L > E$ . Then states  $s$  and  $s''$  differ only in the energy buffer levels. Hence  $H^*(q^1, \dots, q^n, E) \geq H^*(q^1, \dots, q^n, E^L)$ . This proposition provides us a method which is useful for clustering states.

**Remark 5.** *The monotonicity property of the differential value function  $H^*$  provides a justification to group nearby states to form an aggregate state. The value function of the aggregated state will be the average of the value function of the states in a partition. If the difference between values of states in a cluster is not much, the value function of aggregated state will be close to the value function of the unaggregated state. Thus, the policy obtained from the aggregated value function is likely to be close to the policy obtained from unaggregated states. Without the monotonicity property, states may be grouped arbitrarily and consequently, state aggregation may not yield a good policy.*

**Remark 6.** *In the case of MDP with large state-action space, one goes for function approximation based methods (see Chapter 8 in [37]). However, if one combines Q-learning with function approximation, we do not have convergence guarantees to the optimal policy unlike Q-learning without function approximation (Q-learning with tabular representation [37]). However, when Q-learning is combined with state-aggregation (QL-SA) we continue to have convergence guarantees (see Section 6.7 in [5]). Q-learning using state aggregation can produce good policies only when the value function has a monotonicity structure, which is proved in the previous remark.*

### 5.1.1 Clustering

The data and energy buffers are quantized and using this we formulate the aggregate state-action space. The quantization of buffer space is described next. We predefine data buffer and energy buffer partitions (or quantization levels)  $d_1, d_2, \dots, d_s$  and  $e_1, e_2, \dots, e_r$  respectively. The partition (or quantization level)  $d_i, (i \in \{1, \dots, s\})$  corresponds to a given range  $(x_L^i, x_U^i)$  and is fixed, where  $x_L^i$  and  $x_U^i$  represent the prescribed lower and upper data buffer level limits. In a similar manner the quantization level  $e_j, (j \in \{1, \dots, r\})$  (or energy buffer partition) corresponds to a given interval  $(y_L^j, y_U^j)$ , where  $y_L^j$  and  $y_U^j$  represent the prescribed lower and upper energy buffer level limits. As an example, suppose  $D_{MAX} = E_{MAX} = 10$  and each of the buffers are quantized into three levels, i.e.,  $s = r = 3$ . An instance of data and energy buffer partition ranges in this scenario can be  $y_L^1 = x_L^1 = 0, y_U^1 = x_U^1 = 3, y_L^2 = x_L^2 = 4, x_U^2 = y_U^2 = 7, x_L^3 = y_L^3 = 8, y_U^3 = x_U^3 = 10$ . Here Partition 1 corresponds to the number of data (energy) bits (units) in the range  $(0, 3)$ , while Partition 3 corresponds to the number of data (energy) bits (units) in the range  $(8, 10)$ . The following inequalities hold with respect to the partition limits:

$$0 = x_L^1 < x_U^1 < x_L^2 < x_U^2 < \dots < x_L^s < x_U^s = D_{MAX} \text{ and} \\ x_U^i + 1 = x_L^{i+1}, \quad 1 \leq i \leq s - 1.$$

Similarly,

$$0 = y_L^1 < y_U^1 < y_L^2 < y_U^2 < \dots < y_L^r < y_U^r = E_{MAX} \text{ and} \\ y_U^i + 1 = y_L^{i+1}, \quad 1 \leq i \leq r - 1.$$

### 5.1.2 Aggregate States and Actions

We define an aggregate state as  $s' = \{l^1, \dots, l^{n+1}\}$ , where for  $1 \leq i \leq n$ ,  $l^i$  is the data buffer level for the  $i^{th}$  node and  $l^{n+1}$  is the energy buffer level. So  $l^i \in \{1, \dots, s\}$ ,  $1 \leq i \leq n$  and  $l^{n+1} \in \{1, \dots, r\}$ . An aggregate action corresponding to the state  $s'$  is an  $n$ -tuple  $t'$  of the form  $t' = (t^1, \dots, t^n)$ , where  $t^i \in \{1, \dots, l^{n+1}\}$ ,  $1 \leq i \leq n$ . Each component in  $t'$  indicates an energy level. By considering the data level in all the nodes, the controller decides on an energy level for each node. Thus the energy level indicates the energy partition which can be supplied to the node. For instance, if  $D_{MAX} = E_{MAX} = 15$ ,  $s = r = 3$  and there are two nodes in the system, then an example aggregate state is  $s' = (1, 1, 3)$ . Suppose the controller selects the aggregate action  $t' = (2, 1)$ , which means that the controller decides to give  $u$  number of energy bits to Node 1, and  $v$  number of energy bits to Node 2, with  $y_L^2 \leq u \leq y_U^2$  and  $y_L^1 \leq v \leq y_U^1$ , respectively.

### 5.1.3 Cardinality Reduction

Note that  $s \ll D_{MAX}$ ,  $r \ll E_{MAX}$ . Let the aggregated state and action spaces be denoted by  $S'$  and  $A'$  respectively. The aggregated state-action space has cardinality  $|S' \times A'|$ . Thus, the cardinality of the state-action space is reduced to a great extent by aggregation. For instance,

in the case of four nodes sharing energy from one EH source and  $E_{MAX} = D_{MAX} = 30$ , the cardinality of the state-action space without state-aggregation is  $|S \times A| \approx 30^9$ . However, with four partitions each for the data and energy buffers, the cardinality of the state-action space with aggregation is  $|S' \times A'| \approx 4^9$ .

## 5.2 Approximate Learning Algorithm

We now explain our approximate learning algorithm for the energy sharing problem. It is based on Q-learning and state aggregation. Although the straightforward Q-learning algorithm described in Section 4 requires complete state information and is not computationally efficient with respect to large state-action spaces, its state-aggregation based counterpart requires significantly less computation and memory space. Also our experiments show that we do not compromise much on the policy obtained either (see Fig. 9b).

### 5.2.1 Method

Let  $s' = \{l_k^1, \dots, l_k^{n+1}\}$  be the aggregate state at decision time  $k$ . The action taken in  $s'$  is  $t' = (t_k^1, \dots, t_k^n)$ . The Q-value  $Q(s', t')$  indicates how good an aggregate state-action tuple is. The algorithm proceeds with the following update rule:

$$Q_{k+1}(s', t') = (1 - \alpha(k))Q_k(s', t') + \alpha(k) \left( c(s', t') + \min_{b \in A'(j')} Q_k(j', b) - \min_{u \in A'(r')} Q_k(r', u) \right), \quad (30)$$

where  $j'$  is the aggregate state obtained by simulating action  $t'$  in state  $s'$ . Also,  $r'$  is a reference state and  $\alpha(k)$ ,  $k \geq 0$  is a positive step-size schedule satisfying the conditions mentioned in Section 4.2. To facilitate exploration, we employ the mechanisms described in Section 4.2. Convergence of Q-learning with state aggregation is discussed in Section 6.7 of [5].

**Remark 7.** *The aggregate state in every step of the iteration (30) is computed by knowing the amount of data present in each sensor node. A viable implementation would just need a mapping of the buffer levels to these partitions, using which the controller can compute the aggregate state for any combination of buffer levels. Since this method requires storing of Q-value of the aggregate state-action pair and  $|S' \times A'| \ll |S \times A|$ , the number of Q-values stored is much less compared to the unaggregated Q-learning algorithm. The computational complexity of the method described above is dependent on the size of the aggregate state-action space and the number of iterations required to converge to an optimal policy (w.r.t the aggregate state-action space). For instance, in the case of four sensor nodes, the size of the state-action space grows to  $\approx 30^9$  with the data and energy buffer sizes being 30 each. The number of iterations that the above method requires to find a near-optimal policy is  $10^9$  with six partitions of the buffer size as compared to Q-learning without state aggregation (Section 4.2) which requires at least  $10^{11}$  iterations.*

**Remark 8.** *It must be observed that using (30), the controller decides the partition and not the number of energy bits to be distributed, i.e., it finds an optimal aggregate action for*

every aggregate state. It follows from this that, in order to find the aggregate action for an aggregate state, the knowledge of the exact buffer levels is not required (since this is based on the  $Q$ -values of aggregate state-action pairs). In this manner (30) is beneficial. The optimal policy obtained using (30) would indicate only the energy levels. An added advantage of the above approximation algorithm is that the cost structure discussed in Section 3 holds good here as well.

### 5.2.2 Energy distribution

Note that once an aggregate action is chosen for a state, the energy division is random adhering to the action levels chosen. For instance, let's assume that there are two sensor nodes in the system. Data and energy buffers have three partitions each and thus  $s = 3$ ,  $r = 3$ . Here  $y_L^1 = 0$  and  $y_U^3 = E_{MAX}$ . Suppose the number of energy bits in the energy buffer is  $z$  and those bits belong to partition 3. Let the number of data bits at nodes 1 and 2 be  $x$  and  $y$ , respectively. Here  $x$  and  $y$  belong to partition 2. Hence the aggregate state is  $(2, 2, 3)$ . The controller decides on the aggregate action  $(1, 2)$ . Thus  $x_L^1$  bits of energy is provided to Node 1, while Node 2 is given  $x_L^2$  bits of energy. The remaining number of bits in the buffer will be  $r = z - (x_L^1 + x_L^2)$ . In order to distribute these bits, the proportions of data  $p_1 = \frac{x}{x+y}$  and  $1 - p_1 = \frac{y}{x+y}$  are computed. Each of the  $r$  bits are supplied to Node 1 with probability  $p_1$  and to Node 2 with probability  $1 - p_1$ . If  $u$  and  $v$  represent the total number of energy bits provided to Nodes 1 and 2 respectively, then  $u \leq x_U^1$ ,  $v \leq x_U^2$  and  $(u - x_L^1) + (v - x_L^2) \leq r$ . It must be observed that even though an aggregate action chosen requires knowledge of only the aggregate state, the random distribution of energy (after a control is selected using (30)), is achieved by knowing the exact buffer levels.

**Remark 9.** *An advantage of using state-aggregation with  $Q$ -learning is that it has convergence guarantees (Chapter 6, Section 6.2 [5]). This overcomes the problem of basis selection for function approximation in the case of large state-action spaces. We have tried different partitioning schemes manually and all the schemes resulted in close policy performance. Also, we observed that increasing the number of partitions improves the policy performance (see Fig. 6 in Section 6) .*

## 5.3 Cross Entropy using State Aggregation and Policy Parameterization

The cross-entropy method is an iterative approach ([35]) that we apply to find near-optimal stationary randomized policies for the energy sharing problem. The algorithm searches for a policy in the space of all stationary randomized policies in a systematic manner. We define a class of randomized stationary policies  $\{\pi^\theta, \theta \in \mathbb{R}^M\}$ , parameterized by a vector  $\theta$ . For each pair  $(s, a) \in S' \times A'$ ,  $\pi^\theta(s, a)$  denotes the probability of taking action  $a$  when the state  $s$  is encountered under the policy corresponding to  $\theta$ . In order to follow the cross entropy approach and obtain the optimal  $\theta^* \in \mathbb{R}^M$ , we treat each component  $\theta_i$ ,  $i \in \{1, 2, \dots, M\}$  of  $\theta$  as a normal random variable with mean  $\mu_i$  and variance  $\sigma_i$ . We will refer to these two

quantities (the parameters of the normal distribution) as *meta-parameters*. We will tune the meta-parameters using the cross-entropy update rule (32) to find the best values of  $\mu_i$  and  $\sigma_i$  which will correspond to a mean of  $\theta_i^*$  and a variance of zero. The cross entropy method works as follows: Multiple samples of  $\theta$  namely  $\theta^1, \theta^2, \dots, \theta^N$  are generated according to the normal distribution with the current estimate of the meta-parameters. Each sampled  $\theta$  will then correspond to a stationary randomized policy. We compute the average cost  $\lambda(\theta)$  of an SRP determined by a sample  $\theta$  by running a simulation trajectory with the policy parameter fixed with the sample  $\theta$ . We perform this average cost computation for all the sampled  $\theta^i, i \in \{1, 2, \dots, N\}$ , i.e., we compute  $\lambda(\theta^1), \lambda(\theta^2), \dots, \lambda(\theta^N)$ . We then update the current estimates of the meta-parameters based on only those sampled  $\theta$ 's (policies) whose average cost is lower than a threshold level (see (32)).

**Remark 10.** *The Cross Entropy method is an adaptive importance sampling [9] technique. The specific distribution from which the parameter  $\theta$  is sampled is known as the importance sampling distribution. The Gaussian distribution used as the importance sampling distribution yields analytical updation formulas (32) for the mean and variance parameters (see [21]). For this reason, it is convenient to use the Gaussian vectors for the policy parameters.*

### 5.3.1 Policy Parameterization

Let  $\lambda(\theta)$  be the average cost of the system when parameterized by  $\theta = (\theta_1, \dots, \theta_M)^\top$ . An optimal policy  $\theta^*$  minimizes the average cost over all parameterizations. That is,

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^M} \lambda(\theta).$$

An example of parameterized randomized policies, which we use for the experiments (involving state aggregation) in this paper are the parameterized Boltzmann policies having the following form:

$$\pi^\theta(s, a) = \frac{e^{\theta^\top \phi_{sa}}}{\sum_{b \in A(s)} e^{\theta^\top \phi_{sb}}} \quad \forall s \in S', \forall a \in A'(s), \quad (31)$$

where  $\phi_{sa}$  is an  $M$ -dimensional feature vector for the aggregated state-action tuple  $(s, a)$  and  $\phi_{sa} \in \mathbb{R}^M$ . The parameterized Boltzmann policies are often used in approximation techniques ([8, 7, 1, 37, 38]) which deal with randomized policies.

**Remark 11.** *The probability distribution over actions is parameterized by  $\theta$  in the cross entropy method. Since actions in every state need to be explored, the distribution needs to assign a non-zero probability for every action feasible in a state. Hence the probability distribution must be chosen based on these requirements. The Boltzmann distribution for action selection fits these requirements and is a frequently used distribution in the literature (see [37, 38]) on policy learning and approximation algorithms.*

As noted in the beginning of this subsection, the parameters  $\theta_1, \dots, \theta_M$  are samples from the distributions  $N(\mu_i, \sigma_i), 1 \leq i \leq M$ , i.e.,  $\theta_i \sim N(\mu_i, \sigma_i), \forall i$ .

### 5.3.2 Method

Initially  $M$  parameter tuples  $\{(\mu_i^1, \sigma_i^1), 1 \leq i \leq M\}$  for the normal distribution are picked. The policy is approximated using the Boltzmann distribution. The method comprises of two phases. In the first phase trajectories corresponding to sample  $\theta$ s are simulated and the average cost of each policy is computed. The second phase involves updation of the meta parameters. The algorithm proceeds as follows:

Let iteration index  $t$  be set to 1.

*First Phase:*

1. Sample parameters  $\theta^1, \dots, \theta^N$  are drawn independently from the normal distributions  $\{N(\mu_i^t, \sigma_i^t), 1 \leq i \leq M\}$ . For  $1 \leq j \leq N$ ,  $\theta^j \in \mathbb{R}^{M \times 1}$  and  $\theta_i^j$  is sampled from  $N(\mu_i^t, \sigma_i^t)$ .
2. A trajectory is simulated using probability distribution  $\pi^{\theta^j}(s, a)$ ,  $1 \leq j \leq N$ . Hence at every aggregate state  $s$  an aggregate action  $a$  is picked according to  $\pi^{\theta^j}(s, \cdot)$ . Once an aggregate action is chosen for a state, the energy distribution is carried out as described in Section 5.2.2.
3. The average cost per step of trajectory  $j$  is  $\lambda(\theta^j)$  and is computed for the trajectory simulated using  $\theta^j$ . By abuse of notation we denote  $\lambda(\theta^j)$  as  $\lambda_j$ .

*Second Phase:*

4. A quantile value  $\rho \in (0, 1)$  is selected.
5. The average cost values are sorted in descending order. Let  $\lambda_1, \dots, \lambda_N$  be the sorted order. Hence  $\lambda_1 \geq \dots \geq \lambda_N$ .
6. The  $\lceil (1 - \rho)N \rceil^{th}$  average cost is picked as the threshold level. So, let  $\hat{\lambda}_c = \lambda_{\lceil (1 - \rho)N \rceil}$ .
7. The meta-parameters  $\{(\mu_i^t, \sigma_i^t), 1 \leq i \leq M\}$  are updated (refer [24]) in this phase. In iteration  $t$ , the parameters are updated in the second phase in the following manner:

$$\begin{aligned} \mu_i^{(t+1)} &= \frac{\sum_{j=1}^N I_{\{\lambda_j \leq \hat{\lambda}_c\}} \theta_i^j}{\sum_{j=1}^N I_{\{\lambda_j \leq \hat{\lambda}_c\}}}, \\ \sigma_i^{2(t+1)} &= \frac{\sum_{j=1}^N I_{\{\lambda_j \leq \hat{\lambda}_c\}} \left( \theta_i^j - \mu_i^{(t+1)} \right)^2}{\sum_{j=1}^N I_{\{\lambda_j \leq \hat{\lambda}_c\}}}. \end{aligned} \tag{32}$$

8. Set  $t = t + 1$ .



Steps 1-6 are repeated until the variances of the distributions converge to zero. Let  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)^\top$  be the vector of means of the converged distributions. The near-optimal SRP found by the algorithm is  $\hat{\pi}$  where

$$\hat{\pi}(s, a) = \frac{e^{\boldsymbol{\mu}\phi(s,a)}}{\sum_{b \in A(s)} e^{\boldsymbol{\mu}\phi(s,b)}}, \quad \forall s \in S', a \in A'(s).$$

**Remark 12.** *The computational complexity of the cross entropy method is dependent on the number of updations required to arrive at the optimal parameter vector and the dimension of the vector. For instance in the case of four nodes, with data and energy buffer sizes being 30, the cross entropy method requires  $10^3$  sample trajectories for a hyperparameter  $(\mu, \sigma)$  vector of dimension 50. The parameter  $\theta$  is updated over  $10^3$  iterations to arrive at the optimal parameter vector.*

**Remark 13.** *The heuristic cross-entropy algorithm solves hard optimization problems. It is an iterative scheme and requires multiple samples to arrive at the solution. In general one assumes that the parameter  $\theta$  is unknown (non-random variable) and uses actor-critic architecture to obtain locally optimal policy. However, obtaining gradient estimates in actor-critic architecture is hard as it leads to large variance [20]. On the other hand, in our work, we let the parameter  $\theta$  be a random variable and assume probability distribution over  $\theta$  with hyperparameter  $(\mu, \sigma)$  and use cross-entropy method to tune the hyperparameters. Cross entropy method is simple to implement, parallelizable and does not require gradient estimates. To the best of our knowledge, we are the first to combine the cross-entropy with state aggregation and apply it to a real world problem. In [23], the authors sampled from the entire transition probability matrix to calculate the score function and tested on problems with only small state-action space.*

## 6 Simulation Results

In this section we show simulation results for the energy sharing algorithms we described in Sections 4 and 5. For the sake of comparison we implement the greedy heuristic method in the case when the function  $g$  has a non-linear form. Also, we implement Q-learning to learn optimal policies for the case where we consider the sum of the data at all nodes and the available energy as the state. These methods are as follows:

1. Greedy: This method takes as input the level of data  $q_k^i$  at all nodes and supplies the energy based on the requirement. Since  $g(x)$  is the number of data bits that can be sent given  $x$  bits of allocated energy,  $g^{-1}(y)$  gives the amount of energy required to send  $y$  bits of data. Suppose the energy available in the source is  $e_k$  at stage  $k$ . The greedy algorithm then provides  $t_k$  units of energy, where  $t_k = \min\left(e_k, \sum_{i=1}^n g^{-1}(q_k^i)\right)$ . The energy bits are then shared between the nodes based on the proportion of the requirement of the nodes.

2. Combined Nodes Q-learning: The state considered here is the sum of the data at all nodes and the available energy. Let the state space be  $S_c$  and action space be  $A_c$ . So state  $s_k = \left( \sum_{i=1}^n q_k^i, E_k \right)$ . The control specified is  $t_k$  which is the total energy that needs to be distributed between the nodes. In contrast to the action space in Section 3, here the exact split is not decided upon. Instead, this method finds the total optimal energy to be supplied. The algorithm in Section 4.2 is then used to learn the optimal policies for the state-action space described here.

In the above described methods, after an action  $t_k$  is selected, the proportion of data in the nodes is computed. Thus  $p_i = \frac{q_k^i}{\sum_{j=1}^n q_k^j}$ ,  $1 \leq i \leq n$  is computed at time  $k$ , where  $0 \leq p_i \leq 1$  and

$\sum_{i=1}^n p_i = 1$ . Each of the  $t_k$  bits of energy is then shared based on these probabilities. Let  $u_i$  be

the number of bits provided to node  $i$ . Then in the case of the greedy method,  $\sum_{i=1}^n u_i = t_k$ ,

while in the combined nodes Q-learning method,  $\sum_{i=1}^n u_i \leq t_k$ .

## 6.1 Experimental Setup

- The algorithms described in Section 4 are simulated with two nodes and an energy source. We consider the following settings:
  1. For the case of jointly Markov data arrival and Markovian energy arrival processes, we consider energy buffer size of 20 and data buffer size of 10. The data arrivals evolve as:  $X_k = AX_{k-1} + \omega$ , where  $A$  is a fixed  $2 \times 2$  matrix of coefficients and  $\omega = (\omega_1, \omega_2)^\top$  is a  $2 \times 1$  random noise (or disturbance) vector. Here  $A = \begin{pmatrix} 0.2 & 0.3 \\ 0.3 & 0.2 \end{pmatrix}$ . The energy arrival evolves as  $Y_k = bY_{k-1} + \chi$ , where  $\chi$  is also random noise (or disturbance) variable and  $b = 0.5$  is a fixed coefficient. The components in vector  $\omega$  and  $\chi$  are Poisson distributed. In the simulations, we vary the mean of the random noise variable  $\omega_1$ , while means of  $\omega_2, \chi$  are kept constant.
  2. For the case of i.i.d data and energy arrivals the data and energy buffer sizes are fixed at 14.  $X^1, X^2$  and  $Y$  are distributed according to the Poisson distribution. In the simulations, the mean data arrival at node two is fixed while that at node one is varied.
- The algorithms described in Section 5 are simulated with four nodes and an energy source. We consider the following settings:
  1. For the case of jointly Markov data arrival and Markovian energy arrival processes, we consider energy buffer size of 25 and data buffer size of 10. The data arrivals evolve as:  $X_k = AX_{k-1} + \omega$ , where  $A$  is a fixed  $4 \times 4$  matrix of coefficients and  $\omega = (\omega_1, \omega_2, \omega_3, \omega_4)^\top$  is a  $4 \times 1$  random noise (or disturbance) vector. Here

$A = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.1 & 0.1 \\ 0.2 & 0.1 & 0.1 & 0.1 \end{pmatrix}$  The energy arrival evolves as  $Y_k = bY_{k-1} + \chi$ , where  $\chi$  is also random noise (or disturbance) variable and  $b = 0.5$  is a fixed coefficient. The components in vector  $\omega$  and  $\chi$  are Poisson distributed. In the simulations, we vary the mean of the random noise variable  $\omega_1$ , while means of  $\omega_2$ - $\omega_4$ ,  $\chi$  are kept constant. The energy buffer had 4 partitions while the data buffer had 2 partitions.

2. For the case of i.i.d data and energy arrivals, the buffer sizes are taken to be 30 each. The data and energy buffers are clustered into 6 partitions.  $X^1, X^2, X^3, X^4, Y$  are Poisson distributed. In these experiments, the mean data arrivals at nodes 2, 3 and 4 are fixed while the same at node 1 is varied.

For all Q-learning algorithms ( $\epsilon$ -greedy, UCB based and Combined Nodes), stepsize  $\alpha = 0.1$  is used in the updation scheme. For the  $\epsilon$ -greedy method,  $\epsilon = 0.1$  is used for exploration. In the UCB exploration mechanism, the value of  $\beta$  is set to 1. In our experimental simulations, we consider the function  $g(x) = \ln(1 + x)$  for the i.i.d case and  $g(x) = 2\ln(1 + x)$  for the non-i.i.d case.

## 6.2 Results

Figs. 3, 4a, 4b, 5 and 9a show the performance of the algorithms explained in Section 4. The simulations are carried out with two nodes and a single source. Similarly, Figs. 6, 7a and 7b show the performance comparisons of our algorithms explained in Section 5 with other algorithms. The simulations in this case are carried out with four nodes and a single source. In Figs. 3, 6 jointly Markov data arrival and Markovian energy arrival processes are considered and the noise in data and energy arrival at Node 1, i.e.  $\mathbb{E}[\omega_1]$  is varied while that at the other nodes is kept constant. The i.i.d case of data and energy arrivals is considered in Figs. 4a, 4b, 5, 7a and 7b. In these plots, the mean data arrival at Node 1 ( $\mathbb{E}[X^1]$ ) is varied while keeping that at the other node(s) constant. Figs. 3-9b show the normalized long-run average cost of the policies determined by the algorithms along the y-axis. The mean energy arrival is also fixed.

The Q-learning algorithm is designed to learn optimal policies, hence it outperforms other algorithms, as shown in Figs. 3, 4a, 4b, 5 and 9a. The policy learnt by our algorithm does better compared to the greedy policy and the policy obtained from the combined nodes Q-learning method. Note that Q-learning on combined nodes learns the total energy to be distributed and not the exact split. Hence its performance is poor compared to Q-learning on our problem MDP. Thus, sharing energy by considering the total amount of data in all the nodes is not optimal.

Figs. 7a and 7b show the long-run normalized average costs of the policies obtained from the Greedy method and the algorithms described in Section 5. Since our algorithms are model-free, irrespective of the distributions of energy and data arrival (see Figs. 7a and 7b), our algorithms learn the optimal or near-optimal policies. These plots show that our approximation algorithms outperform the greedy and combined nodes Q-learning methods.

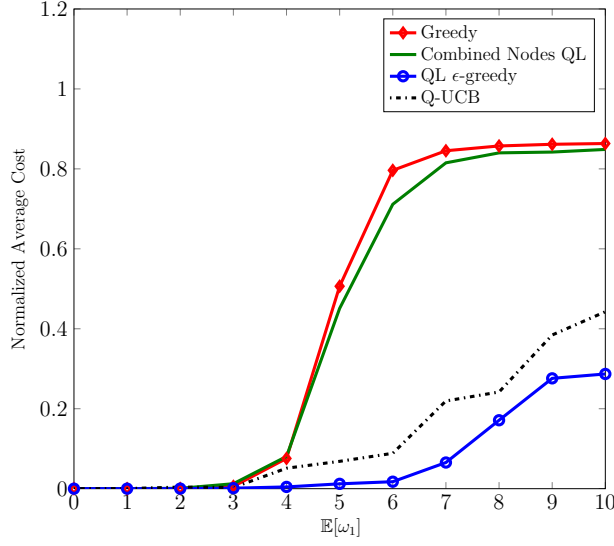
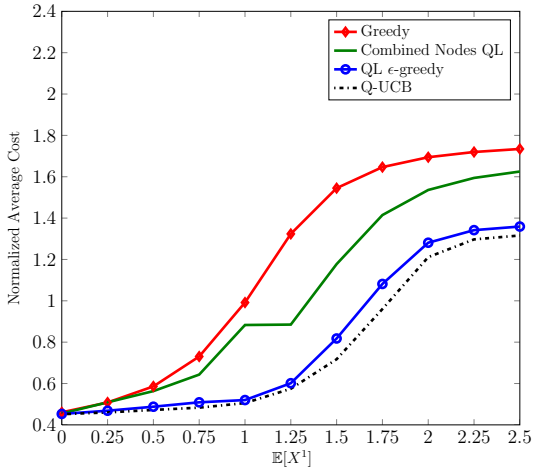
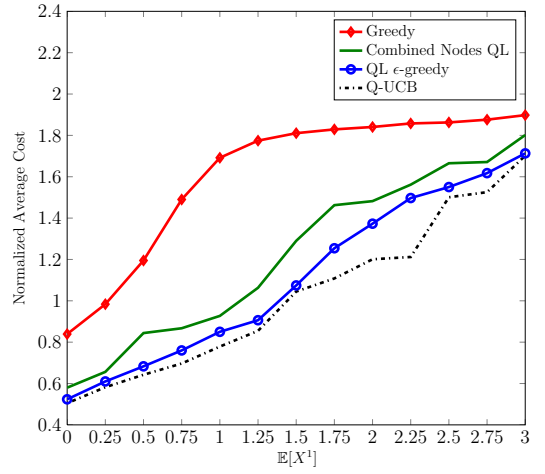


Figure 3:  $E_{MAX} = 20$ ,  $D_{MAX} = 10$ ,  $\omega_1, \omega_2, \chi$  are Poisson distributed with  $\mathbb{E}[\omega_2] = 1.0, \mathbb{E}[\chi] = 20$ ,



(a)  $X^1, X^2, Y$  are Poisson distributed with  $\mathbb{E}[Y] = 13, \mathbb{E}[X^2] = 1.0$



(b)  $X^1$ : Poisson distributed,  $X^2$ : hyperexponential distributed and  $Y$ : Exponential distributed and  $\mathbb{E}[X^2] = 0.625, \mathbb{E}[Y] = 10$

Figure 4: Performance comparison of policies when  $E_{MAX} = D_{MAX} = 14$

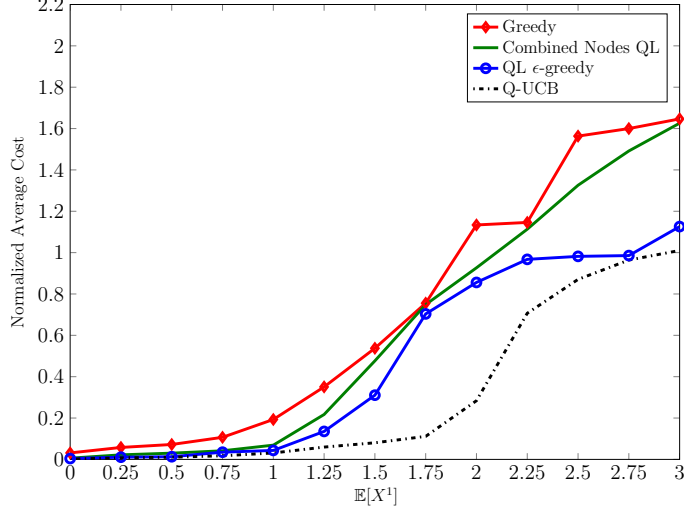


Figure 5: Performance comparison of policies with  $E_{MAX} = D_{MAX} = 30$  when  $X^1, X^2, Y$  are Poisson distributed with  $\mathbb{E}[Y] = 25, \mathbb{E}[X^2] = 1.0$

It can be observed that the gap between the average costs obtained from the combined nodes Q-learning method and the approximate learning algorithm (see Section 5.2) increases with an increase in the number of nodes. This is clear from Figs. 4a and 7a. This occurs because the combined nodes Q-learning method wastes energy and the amount of wastage increases with an increase in the number of nodes.

Fig. 8 shows the variation in average cost with different number of partitions of data and energy buffers used in state aggregation. As the number of partitions increase, the number of clusters also increase resulting in better policies.

The single-stage cost function as defined in (11), includes the effect of action in the conversion function  $g(\cdot)$ . The effect of the action taken can be explicitly included in the single-stage cost function of the following form:

$$c(s_k, T(s_k)) = \sum_{i=1}^n (r_1 * (q_k^i - g(T^i(s_k)))^+ + r_2 * T^i(s_k)), \quad (33)$$

where  $r_1, r_2$  are the tradeoff parameters,  $r_1 + r_2 = 1$  and  $r_1, r_2 \geq 0$ . The above equation is a convex combination of the sum of data queue lengths and the collective energy supplied to the nodes. It can be observed that the single-stage cost function (11) used in our MDP model can be derived from (33) by taking  $r_1 = 1$  and  $r_2 = 0$ . When  $r_1 > 0$  and  $r_2 > 0$ , the cost structure (33) gives importance to the data queue length as well as the amount of energy supplied. The performance comparison of our algorithms (described in Section 4) with the greedy and combined nodes Q-learning methods using this single-stage cost function is shown in Fig. 9a. For the simulations, buffer sizes are fixed at 14 and  $X^1, X^2, Y$  are distributed according to the Poisson distribution with  $\mathbb{E}[Y] = 13$  and  $\mathbb{E}[X^2] = 1.0$ .

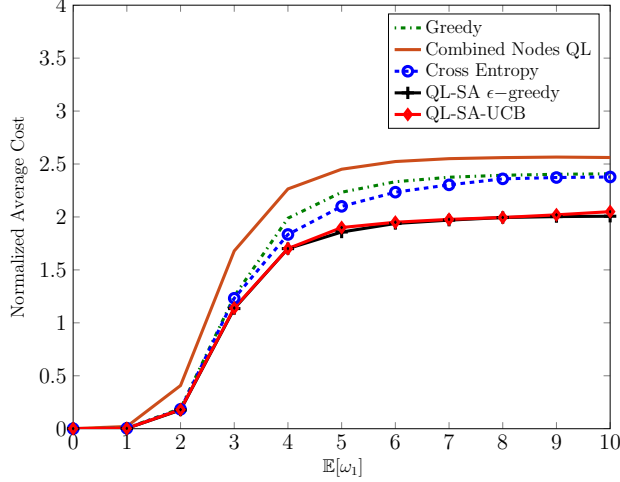
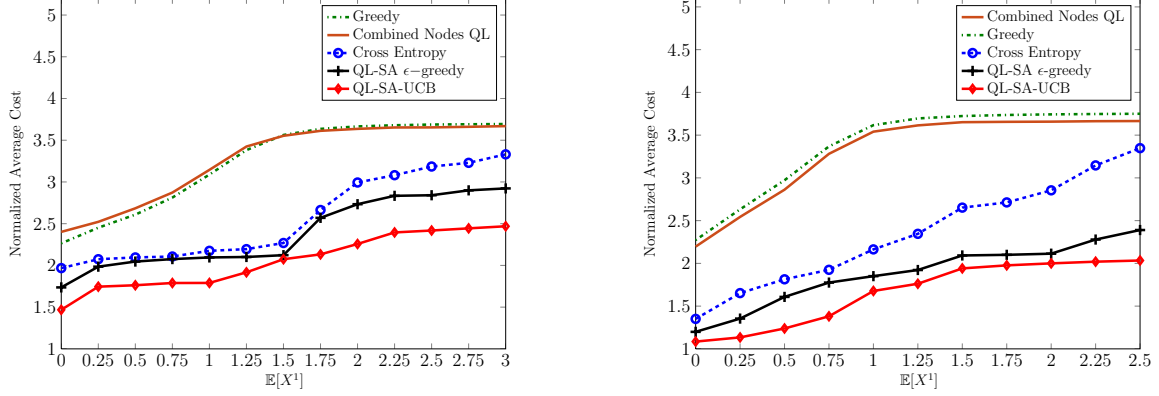


Figure 6:  $E_{MAX} = 25$ ,  $D_{MAX} = 10$ ,  $\omega_1$ - $\omega_5$  are Poisson distributed with  $\mathbb{E}[\omega_2] = \mathbb{E}[\omega_3] = \mathbb{E}[\omega_4] = 1.0$ ,  $\mathbb{E}[\chi] = 5$

In Fig. 9a, the x-axis indicates the change in data rate of Node 1. This setup is akin to that used in Fig. 4a. The y-axis indicates the normalized average queue length of all the nodes. We considered values  $r_1 = 0.7$  and  $r_2 = 0.3$ . The plot indicates only the average queue length of all nodes, since our objective is to minimize the average delay of transmission of data (which is related to the data queue length). From Fig. 9a, it can be observed that all learning algorithms show an increase in the collective average queue length (referred to as the normalized average cost in Figs. 4a-8). This occurs because by using the cost function (33) the learning algorithms (Q-learning with UCB and  $\epsilon$ -greedy exploration as well as combined nodes Q-learning) give less importance to the queue length component in the cost function. Thus the policies learnt by these algorithms minimize the energy usage albeit with an increase in data queue length. As the figure shows, the learning algorithms we described in Section 4 perform much better compared to the greedy and combined nodes methods.

In Fig. 9b, the performance comparison of Q-learning with and without state aggregation is shown for the case of two nodes and an EH source (i.i.d case) and compared with greedy and combined nodes Q-learning method. The  $\epsilon$ -greedy exploration mechanism is used for both algorithms. The experimental setup is similar to that used in Fig. 4a. The x-axis indicates the variation in data rate of Node 1, while the y-axis indicates the normalized average cost of the nodes. The algorithm in Section 5.2 was simulated by partitioning the data and energy buffers into 3 partitions each. It can be observed in Fig. 9b that Q-learning with state aggregation performs better than the greedy and combined nodes methods. However since Q-learning with state aggregation algorithm finds near-optimal policy, its performance is not as good as the algorithm in Section 4.2 with the same exploration mechanism.

**Remark 14.** *The Greedy algorithm distributes the available energy among the sensor nodes*



(a)  $X^1, X^2, X^3, X^4, Y$  are Poisson distributed with  $\mathbb{E}[X^2] = \mathbb{E}[X^3] = \mathbb{E}[X^4] = 1.0$ ,  $\mathbb{E}[Y] = \mathbb{E}[X^3] = \mathbb{E}[X^4] = 0.7$ ,  $X^2$  is hyperexponentially distributed and  $Y$  has the exponential distribution with  $\mathbb{E}[Y] = 20$

Figure 7: Performance comparison of policies when  $E_{MAX} = D_{MAX} = 30$

based on the proportion of data available in the nodes. It shares all the available energy at every decision instant without storing it for future use. We compare our algorithms with the Greedy algorithm in order to show that myopic strategy may not be optimal. Our results show that one has to devise the policy not only for the present requirement for energy but also for the future energy requirements as well. This idea is naturally incorporated in our RL algorithms. Moreover, Greedy policy is optimal when the conversion function  $g$  is linear. This has been derived in [36] for the case of single sensor. The performance of the algorithm proposed in [30] with non-linear  $g$  is compared with the performance of the greedy method. Thus, the comparison of the performance of our algorithms with the greedy method also follows naturally from the earlier cited works.

The Combined Nodes Q-learning method learns the policy which maps the total number of data bits available in all the nodes to the total amount of energy required. The energy sharing between the nodes is then based on the proportion of data available in the nodes. Under the Combined Nodes Q-learning algorithm, the state space is greatly reduced, i.e., instead of the cartesian product of states in each node (as in our Q-learning method with and without state aggregation), it is just the sum of the states of the sensor nodes. So, the learning is faster in combined nodes Q-learning algorithm. However, the policy learnt is suboptimal as was shown in Figs. 4a-9b and performs poorly in comparison with our algorithms. So, we compare our algorithms with Combined nodes Q-learning to illustrate the tradeoff of size of the state space with the nature of the obtained policy.

Note that our RL algorithms learn the energy sharing policy not quantized to a single point but considers energy sharing among the sensor nodes. Learning an optimal energy sharing scheme is a difficult problem. Hence, we would like to understand how well our algorithms perform against a simple heuristic policy such as Greedy or a policy obtained from the Combined nodes Q-learning method.

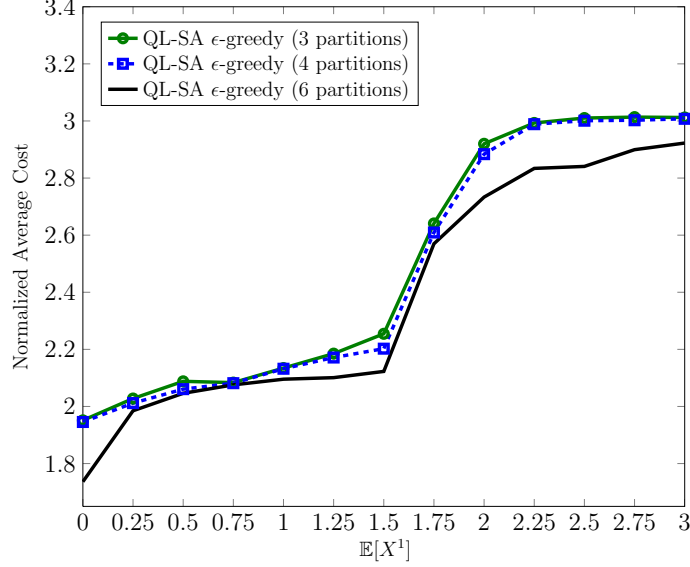


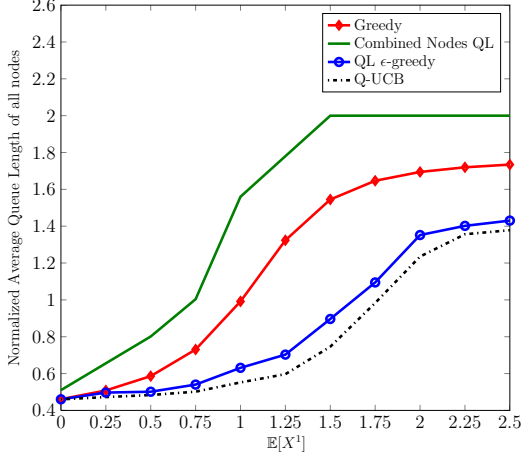
Figure 8: Performance of QL  $\epsilon$ -greedy with different number of data and energy buffer partitions, when  $X^1, X^2, X^3, X^4, Y$  are Poisson distributed with  $\mathbb{E}[X^2] = \mathbb{E}[X^3] = \mathbb{E}[X^4] = 1.0$ ,  $\mathbb{E}[Y] = 25$  and  $D_{MAX} = E_{MAX} = 30$

**Remark 15.** The function  $g(\cdot)$  gives the number of bits that can be transmitted using certain units of energy. Our algorithms work regardless of the forms of  $g$ . RL algorithms use the simulation samples to learn the energy sharing policy by trying out various actions in each of the states. In our problem, at time  $k$  let us assume we are in state  $s_k = (q_k^1, q_k^2, \dots, q_k^n, E_k, X_{k-1}, Y_{k-1})$ , i.e., the data in the data buffer and energy in the energy buffer are fixed to some values. Based on the current  $Q$ -value, we share the energy available to the various sensor nodes by selecting action  $T_k = (T_k^1, T_k^2, \dots, T_k^n)$ . Depending on the action  $T_k$ , the state of the system evolves according to (1)-(2).

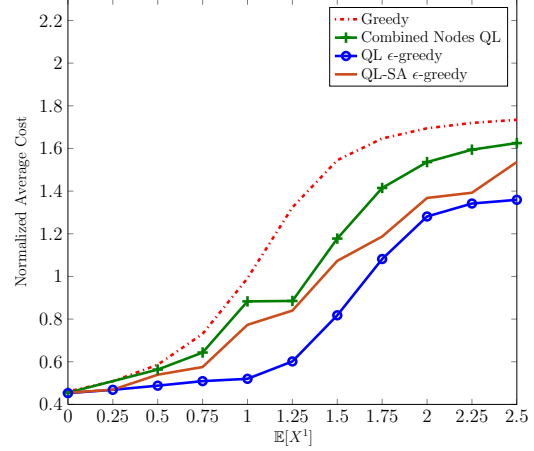
In order to find the next state of the system ((1) - (2)), it suffices to know the number of bits that got transmitted by choosing the action  $T_k$  in slot  $k$  in a real system, which is given by  $g(T_k)$ . It must be noted that we do not need information on the functional form of  $g$  for finding the next state, but only the value of the function for action  $T_k$ . This value can be observed (in a real system) even if we do not have the precise model for the Gaussian function in terms of  $g(\cdot)$ . In other words, all we need is to observe the number of bits that got transmitted by supplying  $T_k$  units of energy.

To update the  $Q$ -value of state-action pair  $(s_k, T_k)$  (see (21)), we need to know the cost  $c(s_k, T_k)$  incurred by choosing action  $T_k$  in state  $s_k$ , which is computed using (11), where again we only require information on  $g(T_k^i)$ ,  $i = 1, 2, \dots, n$ , but not the exact form of  $g(\cdot)$ . Our proposed RL algorithms work by updating  $Q$  values, and such an updation essentially requires the cost information (computed using (11)). Similarly in the cross entropy method, to compute the average cost of the policy, we need to compute the single-stage cost (using (11)). In summary, our algorithms do not require the exact form of  $g(\cdot)$ .





(a)  $r_1 = 0.7, r_2 = 0.3$



(b)  $r_1 = 1$

Figure 9: Performance comparison of policies:  $E_{MAX} = D_{MAX} = 14$  when  $X^1, X^2, Y$  are Poisson distributed with  $\mathbb{E}[Y] = 13, \mathbb{E}[X^2] = 1.0$

*In the case of the greedy algorithm, in order to decide the number of energy units  $T_k$  that need to be shared, the function  $g^{-1}(\cdot)$  and hence the functional form of  $g(\cdot)$  must be known (see Section 6), i.e., one needs to obtain the mathematical model for the conversion function. In comparison, as stated before, our algorithms do not need such information.*

*However, to simulate the environment, we need to know the functional form of the conversion function  $g$ . But, in a real physical system, our algorithms do not require the functional form of  $g$ . Figure 10 illustrates the performance of our algorithms and the Greedy and Combined nodes Q-learning methods for a different form of function  $g(\cdot)$ , i.e.,  $g(\cdot) = \sqrt{3\log(1+x)}$ . The setup is similar to that of Fig. 3. We observe from Fig. 10 that irrespective of the form of  $g(\cdot)$ , our algorithms find good policies, since they do not require this knowledge to do so.*

## 7 Conclusions and Future Work

We studied the problem of energy sharing in sensor networks and proposed a new technique to manage energy available through harvesting. Multiple nodes in the network sense random amounts of data and share the energy harvested by an energy harvesting source. We presented an MDP model for this problem and an algorithm that determines the optimal amount of energy to be supplied to every node at a decision instant. The algorithm minimizes the sum of (data) queue lengths in the data buffers, by finding the optimal energy split profile. In order to deal with the curse of dimensionality, we also proposed approximation algorithms that employ state aggregation effectively to reduce the computational complexity. Numerical experiments showed that our algorithms outperform the algorithms described in Section 6.

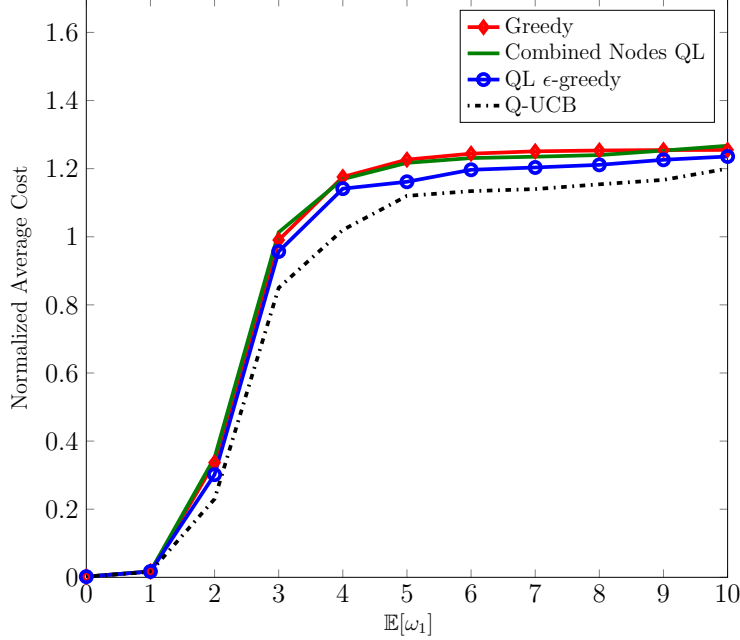


Figure 10:  $E_{MAX} = 20$ ,  $D_{MAX} = 10$ ,  $\omega_1, \omega_2, \chi$  are Poisson distributed with  $\mathbb{E}[\omega_2] = 1.0$ ,  $\mathbb{E}[\chi] = 20$ ,  $g(x) = \sqrt{3 \log(1+x)}$

Our future work would involve applying threshold tuning for state aggregation, gradient based approaches and basis adaptation methods for policy approximation. The partitions formed for clustering the state space (Section 5.1) can be improved by tuning the partition thresholds (see [31]). This method can be employed to obtain improved deterministic policies when state-action space is extremely large. Gradient based methods [6], [20], [8] approximate the policy using parameter  $\theta$  and a set of given (fixed) basis functions  $\{f_k : 1 \leq k \leq n\}$ . Typically a probability distribution over the actions corresponding to a state is defined using  $\theta$  and  $\{f_k\}$ . The parameter is updated using the gradient direction of the policy performance, which is usually the long-run average or discounted cost of the policy. In the approximation algorithm described in Section 5.3, the basis functions used in the policy parameterization are fixed. One could obtain better policies if the basis functions are also optimized. Basis adaptation methods [24], [7] start with a given set of basis functions. The random policy parameter  $\theta$  is updated using simulated trajectories of the MDP on a faster timescale. The basis functions are tuned on a slower timescale. These methods can be employed to find better policies. We shall also develop prototype implementations for this model and test our algorithms.

## Acknowledgements

The authors would like to thank all the three reviewers of [29] for their detailed comments that significantly helped in improving the quality of this report and the manuscript [29]. This work was supported in part through projects from the Defence Research and Development Organisation (DRDO) and the Department of Science and Technology (DST), Government of India.

## References

- [1] Jinane Abounadi, D Bertsekas, and Vivek S Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.
- [2] Anup Aprem, Chandra R Murthy, and Neelesh B Mehta. Transmit power control policies for energy harvesting sensors with retransmissions. *Selected Topics in Signal Processing, IEEE Journal of*, 7(5):895–906, 2013.
- [3] Dimitri P Bertsekas. *Dynamic programming and optimal control, Vol. I*. Athena Scientific Belmont, MA, 1995.
- [4] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 3rd edition, 2007.
- [5] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- [6] S Bhatnagar, H L Prasad, and L A Prashanth. *Stochastic Recursive Algorithms for Optimization*, volume 434 of *Lecture Notes in Control and Information Sciences*. Springer, 2013.
- [7] Shalabh Bhatnagar, Vivek S Borkar, and K J Prabuchandran. Feature search in the grassmanian in online reinforcement learning. *IEEE Journal of Selected Topics in Signal Processing*, 7:746–758, 2013.
- [8] Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.
- [9] James Bucklew. *Introduction to rare event simulation*. Springer, 2004.
- [10] Zhiguo Ding, S.M. Perlaza, I Esnaola, and H.V. Poor. Power allocation strategies in energy harvesting wireless cooperative networks. *Wireless Communications, IEEE Transactions on*, 13(2):846–860, February 2014.

- [11] A El Gamal, J. Mammen, B. Prabhakar, and D. Shah. Optimal throughput-delay scaling in wireless networks - part i: the fluid model. *Information Theory, IEEE Transactions on*, 52(6):2568–2592, June 2006.
- [12] Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *J. Mach. Learn. Res.*, 5:1–25, December 2004.
- [13] Nicolas Gay and W Fischer. Ultra-low-power rfid-based sensor mote. In *Sensors, 2010 IEEE*, pages 1293–1298. IEEE, 2010.
- [14] Munish Goyal, Anurag Kumar, and Vinod Sharma. Power constrained and delay optimal policies for scheduling transmission over a fading channel. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1, pages 311–320. IEEE, 2003.
- [15] B. Gurakan, O. Ozel, Jing Yang, and S. Ulukus. Energy cooperation in energy harvesting communications. *Communications, IEEE Transactions on*, 61(12):4884–4898, December 2013.
- [16] Chin Keong Ho and Rui Zhang. Optimal energy allocation for wireless communications powered by energy harvesters. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 2368–2372. IEEE, 2010.
- [17] Aman Kansal, Jason Hsu, Sadaf Zahedi, and Mani B Srivastava. Power management in energy harvesting sensor networks. *ACM Transactions on Embedded Computing Systems (TECS)*, 6(4):32, 2007.
- [18] Aman Kansal and Mani B Srivastava. An environmental energy harvesting framework for sensor networks. In *Low Power Electronics and Design, 2003. ISLPED’03. Proceedings of the 2003 International Symposium on*, pages 481–486. IEEE, 2003.
- [19] Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, pages 996–1002, 1999.
- [20] Vijay R Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.
- [21] Dirk P Kroese, Sergey Porotsky, and Reuven Y Rubinstein. The cross-entropy method for continuous multi-extremal optimization. *Methodology and Computing in Applied Probability*, 8(3):383–407, 2006.
- [22] Anusha Lalitha, Santanu Mondal, Vinod Sharma, et al. Power-optimal scheduling for a green base station with delay constraints. In *Communications (NCC), 2013 National Conference on*, pages 1–5. IEEE, 2013.

- [23] Shie Mannor, Reuven Y Rubinstein, and Yohai Gat. The cross entropy method for fast policy search. In *ICML*, pages 512–519, 2003.
- [24] Ishai Menache, Shie Mannor, and Nahum Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 134(1):215–238, 2005.
- [25] Dusit Niyato, Ekram Hossain, Mohammad M Rashid, and Vijay K Bhargava. Wireless sensor networks with energy harvesting technologies: a game-theoretic approach to optimal energy management. *Wireless Communications, IEEE*, 14(4):90–96, 2007.
- [26] O. Ozel, K. Tutuncuoglu, Jing Yang, Sennur Ulukus, and A Yener. Transmission with energy harvesting nodes in fading wireless channels: Optimal policies. *Selected Areas in Communications, IEEE Journal on*, 29(8):1732–1743, September 2011.
- [27] Omur Ozel, Kaya Tutuncuoglu, Jing Yang, Sennur Ulukus, and Aylin Yener. Adaptive transmission policies for energy harvesting wireless nodes in fading channels. In *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, pages 1–6. IEEE, 2011.
- [28] Omur Ozel, Jing Yang, and Sennur Ulukus. Optimal broadcast scheduling for an energy harvesting rechargeable transmitter with a finite capacity battery. *Wireless Communications, IEEE Transactions on*, 11(6):2193–2203, 2012.
- [29] Sindhu Padakandla, K J Prabuchandran, and Shalabh Bhatnagar. Energy sharing for multiple sensor nodes with finite buffers. *IEEE Transactions on Communications*, 2015 (Submitted).
- [30] K J Prabuchandran, Sunil Kumar Meena, and Shalabh Bhatnagar. Q-learning based energy management policies for a single sensor node with finite buffer. *Wireless Communications Letters, IEEE*, 2(1):82–85, 2013.
- [31] L.A. Prashanth and S. Bhatnagar. Threshold tuning using stochastic optimization for graded signal control. *Vehicular Technology, IEEE Transactions on*, 61(9):3865–3880, Nov 2012.
- [32] M.L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.
- [33] Zhu Ren, Peng Cheng, Jiming Chen, Ling Shi, and Youxian Sun. Optimal periodic sensor schedule for steady-state estimation under average transmission energy constraint. *Automatic Control, IEEE Transactions on*, 58(12):3265–3271, Dec 2013.
- [34] Zhu Ren, Peng Cheng, Jiming Chen, Ling Shi, and Huanshui Zhang. Dynamic sensor transmission power scheduling for remote state estimation. *Automatica*, 50(4):1235–1242, 2014.

- [35] Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1(2):127–190, 1999.
- [36] Vinod Sharma, Utpal Mukherji, Vinay Joseph, and Shrey Gupta. Optimal energy management policies for energy harvesting sensor nodes. *IEEE Transactions on Wireless Communications*, 9(4):1326–1336, 2010.
- [37] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. MIT Press, 1998.
- [38] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S.A. Solla, T.K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000.
- [39] John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16(3):185–202, 1994.
- [40] John N Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.
- [41] Kaya Tutuncuoglu and Aylin Yener. Short-term throughput maximization for battery limited energy harvesting nodes. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–5. IEEE, 2011.
- [42] Kaya Tutuncuoglu and Aylin Yener. Sum-rate optimal power policies for energy harvesting transmitters in an interference channel. *Communications and Networks, Journal of*, 14(2):151–161, 2012.
- [43] Kaya Tutuncuoglu and Aylin Yener. Cooperative energy harvesting communications with relaying and energy sharing. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE, 2013.
- [44] Jing Yang and Sennur Ulukus. Transmission completion time minimization in an energy harvesting system. In *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, pages 1–6. IEEE, 2010.
- [45] Jing Yang and Sennur Ulukus. Optimal packet scheduling in a multiple access channel with energy harvesting transmitters. *Communications and Networks, Journal of*, 14(2):140–150, April 2012.
- [46] Jing Yang and Sennur Ulukus. Optimal packet scheduling in an energy harvesting communication system. *Communications, IEEE Transactions on*, 60(1):220–230, January 2012.